Reliability Generalization:

Exploring Score Reliability Variance with Ryff's Scale of Psychological Well-Being

Meghan Crouch, BKin

Submitted in partial fulfillment of the requirements for the degree of

Master of Science in Applied Health Sciences

(Kinesiology)

Faculty of Applied Health Sciences, Brock University
St. Catharines, Ontario

Dedication page

To my ‘team’. Who have cheered for me all along, and continue to tell me I am enough. I am so fortunate.

Abstract

The purpose of this study was to conduct a Reliability Generalization (RG; Vacha-Haase, 1998) for Ryff's Scale of Psychological Well-Being (PWB; Ryff, 1989) to characterize the average score reliability, the variability of the score reliability, and explore possible sample and test characteristics that influenced score reliability across studies. Studies were included in the current investigation if they had been published in a peer-reviewed journal, used one or more subscales of the Ryff's PWB, estimated coefficient alpha value(s) for the PWB subscale(s) used, and were written in English. Out of the 924 articles generated by the search strategy, a total of 264 articles were included in the final sample for meta-analysis. The average coefficient alpha for the composite PWB scale was 0.858, with mean coefficient alphas ranging from 0.722 for the Autonomy subscale to 0.801 for the Self-Acceptance subscale. Statistically significant heterogeneity was present across all mean coefficient alphas ($p < .05$), with the heterogeneity index above 95% for both composite and subscale alphas. Consequently, select sample and test characteristics of the primary studies were explored as possible moderator variables on coefficient alpha estimates, with significant differences in score reliability estimates across select demographic and test characteristics. Test length accounted for the majority of variance among alpha coefficients with R^2 values ranging from 40% on the Environmental Mastery subscale to 71% on the Self-Acceptance subscales across the primary studies. In light of the current findings, implications for researchers using Ryff's PWB including informed score reliability reporting practices are discussed.

Keywords: Ryff's Scale of Psychological Well-Being, measurement error, score reliability, reliability generalization, meta-analysis

Acknowledgements

It is the result of the following individual’s efforts, support, and guidance that I was capable of completing this thesis project. I am extremely grateful, and find myself humbled, and at a loss for words (!), to properly express my appreciation.

Dr. Diane Mack

Dr. Philip Wilson

My ‘team’- Mom, Dad, Spencer, Hilary, David, and of course, Christopher

Thank you also to Dr. Matthew Kwan and Dr. Suzie Lane, as well as Elizabeth Yates and Jan Milligan at Brock University Library Services.

Table of Contents

Dedication	ii
Abstract	iii
Acknowledgements	iv
Chapter 1: Literature Review	1
What is Measurement	1
Measurement Error	3
Reliability	4
Factors Affecting Score Reliability	5
The Standards and Other Reporting Recommendations	6
Consequences of Poor Score Reliability	8
Reliability Generalization	9
Ryff's Scale of Psychological Well-Being	11
Research Purpose and Hypotheses	13
Significance of Proposed Research	14
Chapter 2: Methods	15
Sample	15
Coding and Procedures	16
Analysis	17
Chapter 3: Results	19
Search Results and Study Selection	19
PWB Sample and Test Characteristics	19
Mean Reliability and Heterogeneity	21
Moderator Analyses	22
Chapter 4: Discussion	28
Score Reliability Reporting	28

Mean Reliability and Heterogeneity 32
Moderator Analyses 33
Limitations 40
Future Recommendations 43
Conclusions 44
References 46
Appendices 60
Appendix A: Factors Affecting Score Reliability 60
Appendix B: Definitions of Theory-Guided Dimensions of Well-Being 61
Appendix C: Sample Search Strategy 63
Appendix D: RG Coding Form 64
Tables 67
Table 1: Sample and Test Characteristics 67
Table 2: Overall Reliability 71
Table 3: Moderator Analyses for Composite Scale 72
Table 4: Moderator Analyses for Autonomy Subscale 74
Table 5: Moderator Analyses for Environmental Mastery Subscale 76
Table 6: Moderator Analyses for Personal Growth Subscale 78
Table 7: Moderator Analyses for Positive Relations Subscale 80
Table 8: Moderator Analyses for Purpose in Life Subscale 82
Table 9: Moderator Analyses for Self-Acceptance Subscale 84
Figure 80
Figure 1: Flowchart Describing Search Strategy 86

Chapter 1: Literature Review

What is Measurement?

Measurement is a systematic, rule based process in which quantitative values are assigned to represent properties of the individuals, objects, or events (Allen & Yen, 1979; Stevens, 1946). There are three main components of measurement, including the individual, object or event being measured; the instrument selected; and the occasion the measurement occurs (Thye, 2000). While researchers in the physical/ natural sciences use standardized instruments from which measurements can be directly obtained (e.g., weight, concentration, density; Knapp, 1977), researchers examining human behaviour investigate the attributes which characterize such behaviour (Crocker & Algina, 1986). These psychological attributes are theoretical constructs:

> … products of the informed scientific imagination of social scientists who attempt to develop theories for explaining human behavior. The existence of such constructs can never be absolutely confirmed. Thus the degree to which any psychological construct characterizes an individual can only be inferred from observations of his or her behavior. (Crocker & Algina, 1986, p. 4).

Unlike physical attributes, therefore, psychological attributes cannot be directly observed and measured (Crocker & Algina, 1986). In order to measure psychological constructs, operational definitions of the constructs must first be established (Crocker & Algina, 1986; Kline, 2009) which involves the selection of certain observable behaviours (e.g., items) that act as legitimate indicators of the construct of interest. Tests are then developed and utilized to evaluate and score individuals on a sample of these observable behaviours according to a standardized procedure and format (American Educational

Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999, 2014). Additionally, many tests used in the social sciences are an amalgam of items which collectively form the test, while in the physical/natural sciences, measurement may be the result of a single number. Consequently, the processes through which score reliability in the physical/natural sciences and social sciences are determined involve different approaches (Knapp, 1997).

Crocker and Algina (1986) identify five problems inherent in all psychological measurement: (1) no universally accepted method to measure a construct; (2) the relevance and representativeness of a set of items is not exhaustive of the construct being measured; (3) measurement scaling, units and value interpretation is complex and possibly controversial; (4) psychological measurement must relate to measures of other constructs to be useful or have meaning; and (5) all measurement is subject to error.

Despite these recognized difficulties with regards to psychological measurement, Nunnally (1982) succinctly reinforced the importance of measurement quality in the research process:

> Science is concerned with repeatable experiments. If data obtained from experiments are influenced by random errors of measurement, the results are not exactly repeatable. Thus, science is limited by the reliability of measuring instruments and by the reliability with which scientists use them (p. 1589).

As such, measurement theory is “a branch of applied statistics that attempts to describe, categorize, and evaluate the quality of measurements, improve the usefulness, accuracy, and meaningfulness of measurements, and propose methods for developing new and better measurement instruments" (Allen & Yen, 1979, p. 3).

Measurement Error

Regardless of the type of measurement, there is always some degree of error (Crocker & Algina, 1986; Nunnally, 1978; Thompson, 2003). In fact, as early as the 17th century, notable astronomer Galileo recognized errors of observation were evenly distributed, tending to cluster around a single value (Traub, 1997). It was not until 1904, however, when a psychologist by the name of Charles Spearman, laid the foundations of what has now become known as Classical Test Theory (CTT; Traub, 1997). According to CTT, any observed score that is obtained from a measurement is the product of the true score (i.e., one that would be obtained if there were no errors of measurement) and measurement error (Schultz & Whitney, 2005). The true score is a hypothetical value based on the "average of the observed scores obtained over an infinite number of repeated testings with the same test" (Crocker & Algina, 1986, p. 109). Although both systematic and random measurement errors exist and influence score accuracy, systematic errors do so consistently (e.g., a weight scale that is incorrectly calibrated and produces values that are five pounds heavier for every individual). Random measurement errors, however, occur solely due to chance happenings (e.g., incorrect scoring, disturbances/distractions in the testing condition) and may influence scores positively or negatively (Crocker & Algina, 1986). Consequently, random errors result in inconsistent and inaccurate scores. Unlike systematic error that only affects validity, random error affects both validity and reliability (Streiner, 2003).

Classical Test Theory is one way of describing how random errors influence test scores and allows researchers to estimate the relationship between true and observed scores (Crocker & Algina, 1986). This estimation, known as the reliability coefficient, is theoretically defined as "the ratio of true score variance to observed score variance" (Crocker & Algina, 1986, p. 116).

Reliability

Reliability concerns the degree to which measurements are repeatable and stable in and across various contexts and conditions (Nunnally, 1978). Within CTT, reliability estimates the amount of measurement error that contaminates the observed score and therefore, is a property of the scores generated from a particular test and not inherent to the test itself (Crocker & Algina, 1986; Vacha-Haase, 1998). The type of measurement error to be accounted for will determine what type of reliability coefficient to estimate and therefore the particular method to obtain this estimation (Cortina, 1993; Dimitrov, 2002; Schultz & Whitney, 2005). As such, within the CTT framework, there are several methods to estimate reliability including test-retest, interrater, form equivalence and internal consistency (Cortina, 1993; Thompson, 2003). Yet, not all indexes of reliability are appropriate in every situation, nor is there just one type of estimate within each method (Cortina, 1993; Hogan, Benjamin, & Brezinski, 2000; Streiner, 2003). Test-retest reliability estimates measurement error due to changes in examinees (Schultz & Whitney, 2005). Such estimates, reflect the stability of the participants' responses across some time interval (e.g., 4 weeks), and require the same test to be completed on two separate occasions (Thompson, 2003). Dimitrov (2002) recommended that test-retest reliability estimates are most appropriate for evaluating characteristics that are stable over time, such as personality. In testing circumstances which involve the rating of an individual by others, measurement error due to interrater inconsistency is appropriate to investigate and estimate (Schultz & Whitney, 2005). Finally, error due to content sampling can be estimated utilizing either equivalence or internal consistency reliability coefficients (Schultz & Whitney, 2005). Equivalence estimates, both parallel and alternate forms, involve the participant completing two versions of a test and evaluating the item score consistency across test occasions (Schultz & Whitney, 2005). As parallel test forms are difficult to create in practice (Dimitrov, 2002; Sijtsma, 2009), alternate test form equivalence utilizes two similar forms of the test (Dimitrov, 2002; Nunnally, 1978). Internal consistency reliability requires a single administration of a test (Streiner, 2003; Thompson, 2003) with several methods of estimation including split half reliability, various Kuder-Richardson formulas and coefficient alpha (Hogan et al., 2000).

The most commonly reported type of internal consistency reliability is coefficient alpha (Hogan et al., 2000; Kline, 2009), also known as Cronbach's (1951) alpha (Cortina, 1993). Coefficient alpha assesses the extent to which a sample's item scores are interrelated on an instrument (Helms, Henze, Sass, & Mifsud, 2006). Typical values for coefficient alpha range from 0 to 1 (Streiner, 2003) with higher scores representing less error and greater internal consistency (Kline, 2009). For the purposes of the current investigation, the focus will be on internal consistency reliability as estimated by coefficient alpha.

Factors Affecting Score Reliability

Reliability is a dynamic property characteristic of tests scores for a group of examinees rather than the actual test itself (AERA, APA, & NCME, 1999, 2014; Crocker

& Algina, 1986; Vacha-Haase, 1998). Therefore a number of factors affect score reliability including sample heterogeneity, the type of reliability coefficient being estimated and specifics of the test or scale (e.g., test length; see Crocker & Algina, 1986 for a review). Appendix A outlines more specific considerations for score reliability estimation (Symonds, 1928). As such, reliability is not absolute and is largely influenced by variance in the test scores themselves (Streiner, 2003; Thompson, 2003).

Although it is beyond the scope of this paper to go into detail regarding all the factors that may affect score reliability, there are two salient characteristics with respect to coefficient alpha outlined in Appendix A worth noting. Among the factors, coefficient alpha is particularly influenced by the number of test items (Cortina, 1993; Streiner, 2003; Thompson, 2003). Specifically, all else remaining equal, simply increasing the number of test items increases the value of alpha (Cortina, 1993; Streiner, 2003; Thompson, 2003). Item content also affects coefficient alpha with high values (e.g., .99) potentially indicative of content redundancy rather than item homogeneity per se (Streiner, 2003). Consequently, despite its frequency of use, methodologists caution the reliance on coefficient alpha as a reliability estimate without an understanding of the characteristics that affect this statistic (Cortina, 1993; Hogan et al., 2000; Sijtsma, 2009a; Sijtsma, 2009b; Streiner, 2003).

The Standards and other Reporting Recommendations

The first edition of the *Standards for Educational and Psychological Testing* (the *Standards*) was published in 1966, prepared by a joint committee representing AERA, APA, and NCME bodies “to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses" (AERA, APA, & NCME, 2014, p.1). Currently in it's fourth edition, the *Standards* promote rigourous measurement and testing practices, and in effect, provide a set of guidelines for those conducting research (AERA, APA, & NCME, 1999, 2014). Applicable to this particular investigation, twenty standards specific to reliability are identified that test developers and users alike should consider. Specifically, the *Standards* highlight that estimates of test score reliability should be detailed for all scores, outlining the method used to estimate score reliability and any descriptive statistics on the samples for which the estimate applies (see AERA, APA, & NCME, 1999, 2014 for further details and examples). Regardless of scholarly objective, the APA Task Force on Statistical Inference similarly mandates that researchers provide reliability estimates for participants' scores even if the primary purpose of the study is not psychometric (Wilkinson & APA Task Force on Statistical Inference, 1999). As scores, and in effect, score variability is determined by participant data, estimates of reliability will not remain constant across studies. Still, Vacha-Haase, Kogan, and Thompson (2000) contend "too many researchers erroneously assume that their scores will be as reliable as previously reported reliabilities" (p. 511). Such researchers use score reliability estimates from prior investigations. This act, known as reliability induction, (Vacha-Haase et al., 2000) requires sample composition and score variability to be similar to the original study from which the reliability estimates were taken (Vacha-Haase et al., 2000). Thompson and Vacha-Haase (2000) argue that the minimally acceptable practice for reporting score reliability should a researcher choose to induct previous estimates of score reliability, is to provide justification for doing so. Specifically, a direct comparison between their sample and the data from which they are inducting score reliabilities should be provided for both sample characteristics and score standard deviations (Thompson & Vacha-Haase, 2000). Best practice therefore is for researchers to provide score reliability for their own data (Thompson & Vacha-Haase, 2000) as per the *Standards* (AERA, APA, & NCME, 1999, 2014).

Consequences of Poor Score Reliability

Poor score reliability may compromise the validity of the data, that is, the ability of the data to measure the intended construct (Thompson, 2003). Although not sufficient evidence on its own, score reliability is a necessary condition to establish score validity (Sawilowsky, 2000b; Thompson, 2003) and therefore, the inferences and decisional outcomes that can be made based on those scores (Messick, 1995). Additionally, clear evidence of score integrity is crucial in every investigation, as the reliability of the scores directly influences interpretations of statistical significance and effect size (Baugh, 2002; Thompson, 2003; Thompson & Vacha-Haase, 2000). As Vacha-Haase and Thompson (2011) explained, all statistical analyses within the General Linear Model (GLM) are based on the assumption of perfect, or very good score reliabilities. Poor score reliabilities muddy statistical, clinical, and practical significance estimates (Thompson, 2003), because poor score reliability is unaccounted for in the GLM analyses (Vacha-Haase & Thompson, 2011).

What determines the acceptability of score reliability as either poor or acceptable, however, is somewhat contentious and ambiguous. Many authors cite Nunnally's (1978) "Standards of Reliability" when discussing their scores' reliability, yet recommendations for adequate reliability have often been taken out of context (Lance, Butts, & Michels, 2006). Although Nunnally (1978) stated that a modest reliability of 0.70 or higher would likely suffice in preliminary stages of research when there may be time and energy constraints, Nunnally (1978) prefaced this by highlighting “what a satisfactory level of reliability is depends on how a measure is being used” (p. 245). Several measurement experts have agreed with Nunnally’s (1978) sentiments that acceptable reliability is contingent on decisional outcomes that are to be made as a consequence of test score interpretation (Cortina, 1993; Kane, 2011). As such, acceptable score reliability is determined by the tolerance of error in a specific context (Kane, 2011; Wilson, Mack, & Sylvester, 2011). Unfortunately, the notion that there is a particular threshold magnitude, or gold standard, for acceptable score reliability continues to exist, and may lead to a neglect of important contextual factors that should be considered when making decisions regarding acceptable score reliability estimates (Wilson et al., 2011).

Despite the critical foundation of score reliability for validity and GLM statistical analyses, and the ease with which we are able to calculate it (Cunningham, 1986) there are few aspects in the measurement discipline more difficult to comprehend (Cunningham, 1986). Even today, poor language practices (e.g., the *test* is reliable), misuses (e.g., reliability induction), and misunderstandings of reliability and its assessment, continue to persist (Baugh, 2002; Sijtsma & van der Ark, 2015; Thompson, 2003; Vacha-Haase, 1998; Wilson et al., 2011; Yang & Green, 2015).

Reliability Generalization

Vacha-Haase (1998) proposed a method for examining the score reliability of an instrument across test administrations. A specific type of meta-analysis, reliability generalization (RG), evaluates score reliability variance across studies for a particular instrument and the sources of this variance (Vacha-Haase, 1998). An extension of validity generalization (Schmidt & Hunter, 1977), this seminal approach characterizes "(a) the typical reliability of scores for a given test across studies, b) the amount of variability in reliability coefficients for given measures, and c) the sources of variability in reliability coefficients across studies" (Vacha-Haase, 1998, p. 6).

The reliability coefficient is typically the dependent variable in an RG study, with meta-analytic techniques used to integrate score reliabilities from previous test administrations (Rodriguez & Maeda, 2006; Sánchez-Meca, López-López, & López-Pina, 2013). Test (e.g., number of items, response format) and sample (e.g., gender, age) characteristics are the independent variables in an RG study and are selected as potential contributors to score reliability variance across studies (Henson & Thompson, 2002). Past RG studies investigated on average 8.5 (SD = 4.0) characteristics with participant age, gender, ethnicity, and sample size the most frequently used (Vacha-Haase & Thompson, 2011). Such independent variables may provide insight on how and why score reliability fluctuates across studies (Yin & Fan, 2000). In fact, Thompson and Vacha-Haase (2000) contended that the RG methodology is not "monolithic" (p. 187) and is limited only by the researchers' own insightfulness and creativity.

Since Vacha-Haase's (1998) article proposing this meta-analytic technique, RG has been applied to numerous instruments measuring diverse constructs across a wide range of disciplines (Vacha-Haase & Thompson, 2011). Essentially, any test for which reliability estimates are frequently reported can be selected for RG (Henson & Thompson, 2002). Within the psychological sciences, this protocol has been applied to several instruments including: the Beck Depression Inventory (BDI, Yin & Fan, 2000), the Positive Affect and Negative Affect Schedule (PANAS, Leue & Lange, 2011), the

Ways of Coping Scale (WOCS, Rexrode, Petersen, & O'Toole, 2008), and the NEO Personality Inventory (NEO-PI, Caruso, 2000).

RG studies reinforce the notion that reliability is a dynamic property of test scores, and emphasize the potential influence that various sample characteristics and test factors can have on such reliability estimates (Dimitrov, 2002; Vacha-Haase & Thompson, 2011). Vacha-Haase and Thompson (2011) aptly explain the value of RG studies: "…in themselves directly confront chronic misconceptions that tests are reliable. RG studies in and of themselves communicate the important understanding that score reliabilities vary across administrations and are not secreted into test booklets during the test printing process" (p. 164).

Current RG Investigation Instrument: Ryff's Scale of Psychological Well-Being

Prior to the development of Ryff's Scale of Psychological Well-Being (PWB; Ryff, 1989) conceptions of well-being were mainly concerned with subjective well-being, primarily focusing on positive and negative affect and life satisfaction (Ryff & Singer, 2008). Such a focus emphasized "pleasure attainment and pain avoidance" (Ryan & Deci, 2001; p. 141). Instead, Ryff drew from humanistic, existential, developmental, and clinical psychology fields to develop a theoretically driven self-report instrument combining the overlapping themes from these respective fields (Ryff, 1989). The result was the PWB consisting of six subscales: purpose in life; autonomy; personal growth; environmental mastery; positive relations and; self-acceptance (see Appendix B for Ryff's (1989) definitions of the subscales. The subsequent development and refinement of the PWB involved samples of young (n = 133, mean age = 19.53, SD = 1.57), middle-aged (n = 108, mean age = 49.85, SD = 9.35) and older (n = 80, mean age = 74.96, SD =

7.11) adults from the United States selected to explore potential well-being patterns across the lifespan (Ryff, 1989; Ryff, 2014). Ryff's scale has enhanced the understanding of eudaimonic perspectives of well-being given its focus on self-realization and personal potential, rather than affect and life satisfaction consistent with subjective well-being (Ryan & Deci, 2001; Ryff, 1989).

Although a vast array of conceptual and operational definitions of eudaimonic well-being exist (Huta & Waterman, 2014), the utility of Ryff's scale is suggested with its' translation into more than 30 different languages and appearance in more than 150 scientific journals, spanning diverse topics of scientific inquiry (Ryff, 2014). The scale has been used in both large, nationally representative population samples (Abbot et al., 2006; Clarke, Marshall, Ryff, &Wheaton, 2001; Springer, Pudrovska, & Hauser, 2011), as well as smaller, specific sub-population sample groups (Fava et al., 2001; Mack, Wilson, Gunnell, Gilchrist, Kowalski, & Crocker, 2012; Siconolfi et al., 2013). In addition to exploring developmental trajectories, prominent categories of research investigating and/or incorporating PWB include: personality, family experiences, work and other engagements, biological health and clinical/ intervention studies (See Ryff, 2014 for specific details). Since its inception, however, Ryff's scale has also undergone considerable psychometric scrutiny, with more than 35 studies investigating scale reliability and validity issues (Ryff, 2014). One primary contention is regarding scale length; in its original form, Ryff's PWB instrument consisted of 20 items per subscale, for a total of 120 items (Ryff, 1989). To decrease responder burden and to facilitate its inclusion in population health research, several shortened versions of the instrument have since been created including 84, 54, 42, and 18 item versions (Ryff, 2014). Concern regarding psychometric properties for these shorter formats exist; with researchers disputing factorial validity and dimensionality of Ryff's six-factor model (Abbott et al., 2006; Ryff, 2014).

While estimates of score reliability is often reported within a study adopting the PWB (e.g., Abbott et al., 2006; Mack et al., 2012), comparison across studies utilizing this scale has not been examined. With the existence of multiple versions of the PWB and the Scale's use across cultures, it is an ideal candidate for an RG investigation.

Purpose

Guided by the RG approach proposed by Vacha-Haase (1998) to explore issues of score reliability, the purpose of this investigation was to characterize coefficient alpha estimates and its associated variability across primary reports using Ryff's PWB. Additionally, if the amount of variability in the score reliability across studies could not be explained by sampling error alone, a secondary purpose was to explore the sources of the significant variability across studies. An inherent outcome of this RG investigation was an assessment of current reporting practices for score reliability with respect to the PWB.

In line with psychometric and measurement experts (Cortina, 1993; Kline, 2009; Sijtsma, 2009; Thompson, 2003) and the RG literature to date (López-Pina et al., 2015; Vacha-Haase & Thompson, 2011), it was hypothesized that the characteristics of sample age, number of PWB test items, and language of the PWB would explain significant variability in the coefficient alpha estimates across the primary studies. As for the reporting practices of score reliability within primary studies using Ryff's PWB scale, it was expected that the majority of studies would not meet the current recommendations to report score reliability estimates for their own data (AERA, APA, & NCME, 1999, 2014; Wilkinson & APA Task Force on Statistical Inference, 1999) consistent with Vacha-Haase and Thompson's (2011) RG review.

Significance of Proposed Research

This RG study provides insight on the critical, yet often misunderstood, psychometric issue of score reliability (Henson & Thompson, 2002; Vacha-Haase & Thompson, 2011). Utilizing Ryff's PWB, this research highlights the dynamic nature of score reliability, specifically showcasing how score reliability estimates fluctuate as sample and test characteristics vary across test administrations (Henson, 2001; Henson & Thompson, 2002). In emphasizing specific characteristics that affect score reliability, RG, as a meta-analytic approach, provides useful information worthy of consideration for test users when selecting an appropriate measure, and test format, for their particular research and participant sample (Henson, 2001). In emphasizing the variability of score reliability for Ryff's PWB across test administrations, this investigation reinforces the importance of reporting score reliability for the data to be analyzed, consistent with existing recommendations regarding reliability (AERA, APA, & NCME, 1999, 2014; Wilkinson & APA Task Force on Statistical Inference, 1999).

Chapter 2: Methods

This section outlines the search strategy entered into the electronic databases, inclusion criteria, the selection of the independent variables that were coded, the development of the coding manual, as well as the analyses in this particular study. The following procedure was consistent with the RG recommendations provided by Henson and Thompson (2002) in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses Protocols (PRISMA-P) statement (see Moher et al. (2015) for details).

Sample

Peer reviewed articles published or written after the PWB was first introduced in the scientific literature (i.e., 1989), to March 2016 served as the population of interest for data collection. The sample was generated from PsycINFO, ProQuest Nursing and Allied Health, as well as MEDLINE via Web of Science Complete. The following keywords were entered into the search features of each database: ((eudaimonic OR eudaemonic) AND “well-being”) OR ryff OR “scale* of psychological well-being” OR “model of psychological well-being” (see Appendix for an example of an exact search strategy used for one of the databases). It is important to highlight that researchers refer to Ryff’s Scale using alternate titles (e.g., Mack et al. (2012) refer to it as the Scales of Psychological Well-Being (SPWB) while Springer et al. (2011) refer to it as Ryff’s Model of Psychological Well-Being (RPWB)). After consultation with a Brock University liaison librarian, the aforementioned search strategy ensured pertinent articles utilizing Ryff’s Scale, but possibly calling it by a different name, were included in the current investigation.

All possible articles identified as a function of the keyword search per database were vetted against the study inclusion/exclusion criteria. To be eligible for inclusion in the final sample for data analysis, each study that was identified via the keyword search must have (a) been published in a peer-reviewed journal; (b) used at least one of the subscales of Ryff's PWB; (c) reported coefficient alpha value(s) for the PWB subscale(s) used, and (d) be written in the English language. Studies were additionally excluded when articles published reliability in unusable formats (ranges, statements of "greater than" a specified bound (e.g. greater than .70), or a composite value that did not include all six subscales). Duplicates of studies already identified for inclusion were excluded, and care was also taken to remove publications using redundant samples or alpha values inducted from previous studies. Consistent with the definition provided by Vacha-Haase, Kogan, and Thompson (2000), reliabilities were classified as induction in the current study when authors explicitly cited alpha coefficients from prior investigations and used these values "as the sole warrant for presuming the score integrity of entirely new data [their own]" (p. 512).

Coding and Procedures

Based on past RG coding recommendations (Henson & Thompson, 2002; Vacha-Haase & Thompson, 2011) clear and detailed rules were developed to code the primary articles (see Appendix C). Sample, test and reliability characteristics were extracted from the studies and directly recorded as continuous variables. Specifically, sample characteristics that were coded for and used as independent variables in the RG include: sample size, gender (male, female, or both), ethnicity, and health status. Average age was also extracted from the primary studies and was then used to classify the samples into age categories: children (0 to 12 years), adolescents (13 to 18 years), emerging adults (19 to 25 years), adults (26 to 64 years) and older adults (65 years and older). Test characteristics that were coded for include: the number of subscales of the PWB used in the primary study, number of items per subscale, response format, and language of the PWB. Additionally, study design was coded as either non-experimental, quasi-experimental, or randomized control trial.

One coder coded all studies. A second coder independently coded a random sample (at least 30%) of the studies selected for inclusion. Prior to coder's initiating their work with the data and to reduce concerns pertaining to ambiguity during the coding process, Meghan Kathleen Crouch underwent a period of training and familiarization overseen by Diane Elizabeth Mack; an experienced researcher formally trained in meta-analysis. During this training, questions concerning the coding process and procedure were clarified by DEM. Any encountered problems or discrepancies while coding were discussed and consensus reached between coders. Additionally, interrater reliability was calculated by percent agreement for the data coded by MKC and DEM. The interrater reliability was 95.6% for the current investigation.

Analysis

Although many analytic choices for RG exist (Thompson & Vacha-Haase, 2000; Vacha-Haase, Henson & Caruso, 2002) there is no definite, preferred approach (Sánchez-Meca et al., 2013). In the present study, using the software package, Comprehensive Meta-Analysis Version 3.0 (Borenstein, Hedges, Higgins, & Rothstein, 2005), a random effects model was assumed to obtain summary statistics of reliability coefficients. Specifically, effect sizes, 95% Confidence Intervals (*CI*), Q and I^2 statistics were interpreted from the random effect models. While the *Q* statistic provides an assessment of the statistical significance of the variability, the I^2 index assesses the proportion of the total variability in the alpha coefficients due to *true* heterogeneity, that is, due to actual between-studies variability and not just sampling error (within-study variability; Huedo-Medina, Sanchez-Meca, Marin-Martinez, & Botella, 2006). Additionally, moderator analyses were conducted to explore the effects of study characteristics on the variability of the alphas. Select sample (gender and age category) and test (number of test items per subscale or composite, response format, and language of PWB) characteristics were considered in the moderator analyses. All moderators were deemed categorical, therefore analyses of variance to examine their influence on coefficient alpha were used. Meta-regression analyses were used to report the percentage of variance accounted for by each moderator on coefficient alpha variability.

Chapter 3: Results

Search Results and Study Selection

Figure 1 presents a flowchart outlining the selection process of the studies. The search across the three databases yielded a total of 924 studies, out of which 124 were removed as duplicates. Of the remaining 800 studies, an additional 286 were removed as they did not employ Ryff's PWB (*n* = 215) or were written in a language other than English (*n* = 71). Ninety- three studies were also removed as they combined Ryff's PWB with other measures, reused sample data from previously published studies, or could not be attained in the time frame of the data collection period through the interlibrary loan system. Finally, a further 157 studies were excluded as they provided alpha values for their scores in formats that were not practical for use in this meta-analyses (e.g., ranges, alpha values > than, or composites that did not include all six PWB subscales; *n* = 51), inducted coefficients from previous studies (*n* = 29) or gave no alpha reliability at all (*n* = 77). Therefore, the RG meta-analysis included 264 articles that reported an alpha score reliability estimate for their own data. Specifically, 113 articles reported score reliability estimates for the PWB composite while 134 articles provided score reliabilities for PWB subscales. Seventeen articles provided both PWB composite and subscale score reliabilities.

PWB Sample and Test Characteristics

PWB sample and test characteristics of these primary studies along with the corresponding frequencies (*k*) of the coefficient alphas are presented in Table 1. To highlight, both the composite and PWB subscales were most frequently administered to adult samples (*k* = 67; 44.67%; *k* = 300; 41.03% respectively), although age categories did range from children to older adults. Both the PWB composite (*k* = 125; 83.33%) and PWB subscales (*k* = 639; 87.41%) were primarily administered to mixed gendered samples. Ethnicity was less frequently reported in primary studies, but when it was, samples were most frequently of mixed ethnicity (*k* = 58; 38.67% for composite PWB; *k* = 209; 28.60% for PWB subscales). Similarly, sample health status was often not reported for either the PWB composite or PWB subscales (*k* = 127; 84.67%; *k* = 593; 81.12%).

With regards to the test characteristics, the PWB was translated into 19 different languages for use in the primary investigations with English the most frequently reported language version (*k* = 428; 45.58%). Originally designed as a 120-item instrument by Ryff (1989), the number of items in the composite PWB scale ranged from 6 to 86 across retained studies, with the 18-item version the most frequently reported (*k* = 68; 45.33%). Similarly, the PWB subscales varied in length, ranging from 3 items to the original 20-item per subscale version, with the 14-item the most frequently reported (*k* = 211; 28.86%). Variation in response formats was also apparent for the PWB in the primary studies, and ranged from 2- to 8-point Likert scales. The largest proportion of both the composite and PWB subscale coefficients (*k* = 65; 43.33%; *k* = 453; 61.97% respectively), however, were reported with respect to the 6-point format as per Ryff (1989).

Finally, according to PRISMA-P (Moher et al., 2015), researchers conducting a meta-analysis should assess the risk of biases of the individual studies (formerly referred to as study ‘quality’). Despite this intention at the outset, as the data coding proceeded, the author recognized that the method employed to assess the risk of bias (i.e. study design) had no bearing on the quality of the score reliability reporting (the selected outcome of the meta-analysis). For instance, even if the study was a randomized, controlled trial, only one score reliability was reported despite two group scores being compared across time points (Aschbrenner, Greenberg, & Seltzer, 2009; Deane, Marshall, Crowe, White, & Kavanagh, 2015). As such, the author did not believe study design to be a particularly useful indicator of 'quality' in accordance with reliability recommendations in the *Standards*, nor did it provide additional insight with regards to reliability generalization across studies. Evidently, study design was not coded for and included in this section.

Mean Reliability and Heterogeneity

Summary statistics for all the coefficient alphas are presented in Table 2. It is important to highlight that the majority of studies reported more than one score reliability estimate. Specifically, 150 coefficient alphas were published for the composite PWB scale scores in 130 of the primary studies. Seven hundred and thirty-one additional alpha values for PWB subscales were reported in 151 of the primary studies. Of the six PWB subscales, the most frequently employed was the Purpose in Life subscale (k = 137; 18.74%).

The 150 composite score estimates for the PWB scale yielded a mean coefficient alpha of 0.858 (*95% CI* = 0.846 - 0.869) with reported study values ranging between 0.51 and 0.99. With regards to studies reporting PWB subscales, mean coefficient alphas ranged from 0.722 for the PWB Autonomy subscale (*95% CI* = 0.697 - 0.745) to 0.801 for the PWB Self-Acceptance subscale (*95% CI* = 0.780 - 0.819). Statistically significant heterogeneity was present across all mean coefficient alphas (p < .001), with the heterogeneity index (I_2) above 95% for both composite and subscale alphas. As a result, moderator analyses to explain some of the variation of the alphas across studies was warranted.

Moderator Analyses

Select sample (i.e., age category, gender) and test (i.e., number of items, response format and language) characteristics were explored as possible moderator variables on coefficient alpha estimates. Characteristics were only included for moderator analyses, if $k \geq 3$. Although health status and ethnicity were coded for, very few of the primary studies reported this information. As well, the diversity of the physical and psychological conditions reported across the studies (e.g. types of cancers, multiple sclerosis, HIV, mood disorders), rendered running a moderator analysis on sample health somewhat questionable, as no theoretical rationale could be justified for grouping such diverse health conditions to allow for comparisons. With regards to ethnicity, insufficient detail provided in the primary studies prevented subsequent moderator analyses. Finally, all covariates could not be entered into the meta-regression analysis simultaneously due to: 1) the number of moderators examined in the present investigation relative to *k*, and 2) concerns over collinearity most notably between test characteristic variables.

Composite PWB. Results of the moderator analyses for the coefficient alphas across all studies adopting the composite PWB scale are displayed in Table 3. Significant differences in mean coefficient alphas across age categories for the PWB composite were found (Q = 31.43; p < 0.001) with 4% of the variance explained. The highest mean coefficient alpha (α = 0.908) for the older adult age category PWB scores. The Q-test for gender, however, did not reach statistical significance (p > .05), and consequently, was not considered a sample characteristic that significantly influenced the coefficient alpha estimates across studies. In contrast, tests of heterogeneity revealed statistically significant differences for all test characteristics (i.e., number of items, response format, test language) on the variance of the coefficient alpha estimates for the PWB composite. In keeping with psychometric theory, significant differences existed for the average coefficient alphas based on the number of items in the PWB composite test (Q = 148.29; p < 0.001; R^2 = .49), with the highest mean coefficient alpha (α = 0.94) reported for scores from the 84-item version. As for response format, significant differences between mean coefficient alphas were found (Q = 8.61; p = 0.035 R^2 = .04). Finally, the language of the composite PWB was also a significant moderating factor in the mean coefficient alpha variation (Q = 52.46; p < 0.001) with 10% of variance in coefficient alpha scores explained. All I^2 values were above 96%.

PWB Autonomy subscale. Table 4 displays the results of the moderator analyses for the Autonomy score reliability estimates. The test of heterogeneity revealed significant differences between mean coefficient alpha estimates across age categories (Q = 10.39; p = 0.015) with 23% of the variance accounted for. Adolescent samples had the lowest mean score reliability estimates for the Autonomy subscale (α = 0.641). In contrast, no significant differences (p > .05) in average coefficient alpha values by gender were apparent. The number of items on the Autonomy subscale was a moderating factor on mean coefficient alpha estimates (Q = 117.31; p < 0.001; R^2 = .55), with the lowest mean reliability estimates reported for 3-item subscale scores (α = 0.497). Although only the 5, 6, and 7-point response formats were used with respect to the Autonomy Subscale, significant differences in mean coefficient alphas were reported for scores with various response formats (Q = 6.94; p = 0.031; R^2 = .00). Additionally, significant differences in the mean coefficient alphas were found across different languages of the Autonomy subscale (Q = 106.87; p < 0.001) with 5% of the variance accounted for. The Italian translation of the subscale had the highest mean alpha score reliability (α = 0.795) while the Swedish translation had the lowest (α = 0.494). All I^2 values were above 96%.

PWB Environmental Mastery subscale. Results of the moderator analyses for the coefficient alphas across all studies employing the Environmental Mastery subscale are displayed in Table 5. Both age category and gender did not significantly influence mean score reliability (p = 0.288 and p = 0.262 respectively). Tests of heterogeneity revealed that significant differences in mean coefficient alpha estimates were found across all characteristics related to the PWB itself. Specifically, significant differences in score reliability were found depending on the number of items on the Environmental Mastery subscale (Q = 76.66; p < 0.001) and response formats (Q = 14.67; p = 0.002), with scores from the original 6-point version having the highest estimate (α = 0.739). Forty percent of the variance in coefficient alpha was accounted for by the number of PWB items used, whereas 0% was accounted for by response format. Mean score reliability significantly differed across language versions of the Environmental Mastery subscale (Q = 44.53; p < 0.001; R^2 = .06). Both Portuguese and Japanese mean score reliabilities were below 0.600. All I^2 values were above 96%.

PWB Personal Growth subscale. Results of the Q-tests revealed that only the number of items and the language of PWB significantly influenced mean score reliability for the Personal Growth subscale (see Table 6 for details). Specifically, significant differences in mean coefficient alphas were found across the Personal Growth subscale as the number of items varied (Q = 100.02; p < 0.001; R^2 = .53). Mean score reliability was highest on the 14-item subscale version (α = 0.809) while the alphas for the 3-item version scores were the lowest (α = 0.547). With regards to language, significant differences (Q = 35.42; p < 0.001; R^2 = .00) in mean coefficient alphas were reported for scores across different language versions, with alphas ranging from 0.627 on the Swedish translation to 0.786 on the Italian version of the PG subscale. All I^2 values were above 96%.

PWB Positive Relations subscale. Table 7 displays the results of the moderator analyses for the coefficient alphas across all studies employing the Positive Relations subscale. Neither age category nor gender of the samples significantly influenced mean score reliabilities (p > .05). Yet, significant differences in mean coefficient alpha estimates were reported for scores from the Positive Relations subscales depending on the number of items, response format, and language. Specifically, the test of heterogeneity revealed significant differences in score reliability estimates depending on the number of items in the subscale (Q = 168.21; p < 0.001; R^2 = .71). Similarly, response format significantly influenced mean score reliabilities (Q = 13.09; p = 0.004; R^2 = .00). Significant differences in mean coefficient alphas were also found across various languages of the Positive Relations subscale scores (Q = 47.70; p < 0.001) with 9% of variance accounted for. In particular, scores from the Swedish version had the lowest mean score reliability at 0.554. All I^2 values were above 96%.

PWB Purpose in Life subscale. Results of the moderator analyses for all Purpose in Life subscale score reliabilities appear in Table 8. Age category of the sample significantly influenced mean coefficient alpha (Q = 8.19; p = 0.042; R^2 = .02). Tests of heterogeneity also revealed significant differences in average score reliabilities for the Purpose in Life subscale depending on the number of items (Q = 99.31; p < 0.001; R^2 = .62), response format (Q = 8.72; p = 0.033; R^2 = .00), and language (Q = 542.64; p < 0.001; R^2 = .17) of the subscale. For instance, mean coefficient alphas ranged from 0.418 for the 3-item version to 0.841 on the 14-item version. Similarly, significant differences in the average score reliabilities were found across various translations of the Purpose in Life subscale. The Swedish translation produced scores with an average alpha estimate of 0.254, while the Italian version had an average score reliability of 0.813. All I^2 values were above 98%.

PWB Self-Acceptance subscale. Table 9 displays results of the moderator analyses for all the Self-Acceptance score reliabilities. Tests of heterogeneity revealed statistically significant differences in average score reliabilities across all moderators with the exception of gender. Age category of the sample significantly influenced mean coefficient alpha (Q = 8.66; p = 0.034; R^2 = .26), with scores from the emerging adults age category having the highest average alpha (α = 0.837). While no significant differences (p > .05) in average coefficient alpha values by gender were apparent, there were significant differences in mean score reliabilities for the Self-Acceptance subscale when the number of items, response format, and language varied. Specifically, the number of items on the Self-Acceptance subscale was a moderating factor on mean coefficient alpha estimates (Q = 131.57; p < 0.001) with 54% of the variance in coefficient alpha accounted for. The highest mean reliability estimates reported were for 14-item subscale scores (α = 0.870). With regards to response format for the Self-Acceptance subscale, the 4, 5, 6, and 7-point response formats were used, with significant differences in mean coefficient alphas for scores with various response formats (*Q* = 6.94; $p = 0.031$; $R^2 = .00$). The 6-point response format had the highest mean coefficient alpha ($\alpha = 0.810$). Additionally, significant differences in the mean coefficient alphas were found across different languages for the Self-Acceptance subscale scores (*Q* = 81.61; $p < 0.001$; $R^2 = .10$). The Italian translation of the subscale had the highest mean alpha score reliability ($\alpha = 0.875$) while the Portuguese translation had the lowest ($\alpha = 0.662$). All I^2 values were above 97%.

Chapter 4: Discussion

Reliability is a dynamic property of test scores for a particular group of examinees on a specific administration of a test. To reinforce this notion, and directly challenge reliability myths, Vacha-Haase (1998) proposed RG. As per Vacha-Haase's (1998) approach, the purpose of the current study was to explore score reliability, in the form of coefficient alpha, across primary studies that employed Ryff's Psychological Well-Being Scale. Specifically, score reliability estimates from 264 primary studies were integrated through meta-analytic techniques to characterize the average score reliability, the variability of the score reliability reported across these studies, and the possible sample and test characteristics that explained this variability.

Score Reliability Reporting

Although reporting standards and measurement experts (e.g., Thompson, 1994; Vacha-Haase, 1998) alike have made clear the importance of estimating score reliability for all data to be interpreted in research, seventy-seven (20.81%) of the primary studies identified via the search strategy in the current investigation could not be included in the meta-analysis as researchers did not report coefficient alphas for their data. An additional 29 primary studies (7.84%) were also excluded as they induced reliability coefficients from previous reports. According to Thompson and Vacha-Haase (2000), should researchers choose to induct score reliability from a previous study, the minimally acceptable practice is direct and explicit comparison of both the sample characteristics and standard deviations of the scores for the participants between the studies. Yet, none of the 29 studies that did induct reliability in the current investigation engaged in this minimally acceptable practice. The fact that the authors of these 29 published articles inducted reliability without direct comparison, with an additional 77 failing to report score reliability at all, suggests that score reliability recommendations have yet to be adopted by all researchers in practice. There still appears to be a lack of understanding of the importance of reporting score reliability for the data in hand, and how factors related to the specific sample and test used in any investigation may influence score reliability estimates.

Yet, in comparing the current findings to previous research investigating score reliability reporting practices across a number of behavioural science fields and instruments, the percentage of Ryff's PWB studies (71.35%) with usable reliability information was relatively high. As noted by Green, Chen, Helms, and Henze (2011), there is very little consistency and a great deal of oversight with regards to reliability reporting practices even today. For instance, in Vacha-Haase and Thompson's (2011) review of the RG literature, they reported that the majority (54.6%) of the primary studies did not cite score reliability for their own scores. Similarly, Green et al. (2011) reported that only 28% of published articles provided reliability coefficients for participant data in their review of reliability reporting practices in *Psychological Assessment*, while an additional 11% of articles used previously reported reliability coefficients. More recently, Barry, Chaney, Piazza-Gardner, and Chavarria (2014) reviewed reporting practices in seven of the most prominent journals in the fields of health education and behaviour and reported that only 409 out of the 967 articles (42.3%) reviewed provided reliability estimates for their samples. Finally, in a recent RG investigation of the Yale-Brown Obsessive Compulsive Scale (Y-BOCS), López-Pina et al. (2015) discovered that only 6% of the 2,179 studies employing the Y-BOCS reported score reliability for sample scores, while the remaining 94% of articles merely inducted reliability from other studies! Although these respective findings may not generalize across all fields of scientific inquiry, it appears there is still ample room for improvement of score reliability reporting practices in accordance with recommendations (AERA, APA, & NCME, 2014) and expert opinion (Helms et al., 2006; Thompson & Vacha-Haase, 2000).

Furthermore, although the majority of coefficient alphas were reported for the specific PWB subscales (k = 731; 82.97%), still 17.03% of the coefficient alphas (k = 150) were published for composite PWB scale scores in the current investigation. As many measurement experts have emphasized (Cortina,1993; Schmitt, 1996) although coefficient alpha is not a measure of unidimensionality, one of the assumptions for its appropriate use is essential tau-equivalence; in that a single factor underlies all items with a common factor loading (Green & Yang, 2015). As such, Helms et al. (2006) highlighted that researchers should consider whether the scale's conceptual and structural properties suit the calculation of coefficient alpha a priori. Although Ryff's original six-factor model for well-being has indeed received criticism (e.g., Abbott et al., 2010; Kafka & Kozma, 2002; Springer, Hauser, & Freese, 2006), these studies still support a multidimensional factor structure for well-being, and not a unidimensional model. As such, it is inappropriate to estimate a composite PWB score coefficient alpha using all scale items, as it may violate the tau-equivalence assumption of unidimensionality (see Helms et al. (2006) for a detailed discussion on proper methods to calculate composite reliability and how it differs from total-scale reliability coefficients).

Although calculating composite reliability using all item responses is indeed improper implementation of coefficient alpha (Gignac, 2013; Sijtsma & Emons, 2011), it is a strategy commonly used to replace low subscale alphas (Helms et al., 2006). This was evident in the current investigation, as some authors reported that composite PWB score reliabilities were used in lieu of the lower subscale coefficient alphas, in order to achieve a higher score reliability (e.g. Boylan & Ryff, 2015; Franz et al., 2012; Joshanloo & Ghaedi, 2009). Schmitt (1996) criticized that such a preoccupation to obtain a certain 'acceptable' level of alpha may prevent critical considerations regarding what influences alpha (i.e., the number of items (Cortina, 1993), sample heterogeneity (Thompson, 1994) and issues related to the actual measures construct validity. Similarly, according to Helms (2006): "using alpha should not be pro forma but rather should reflect informed decision making about which set of measurement assumptions one's data best fit" (p. 636). Should researchers choose to use a composite measure of PWB, measurement experts suggest alternative, more appropriate methods for calculating score reliability for multidimensional scales as opposed to coefficient alpha (see Gignac, 2013; Sijtsma & Emons, 2011 for alternative methods). Using appropriate methods for calculating score reliability based on the underlying structure of the test is recommended in the *Standards* (see Standard 2.5, AERA, APA, & NCME, 2014), yet based on the number of coefficient alphas calculated for the composite PWB (k = 150), it appears such considerations are still overlooked by some researchers in practice.

Further, the *Standards* explicitly states that it is insufficient to report total score reliabilities should the subscale scores be interpreted and utilized in the analyses (see Standard 2.3, AERA, APA, & NCME, 2014). Despite this, some authors (e.g., Aschbrenner, Greenberg, & Seltzer, 2009; Vescovelli, Albieri, & Ruini, 2014) only reported composite score alphas even though they interpreted the subscale scores in their investigations. Ironically, although total score reliabilities may be sufficiently high, the implications of the poor subscale score reliabilities are not eliminated. Poor score reliability may compromise construct validity, as well as study significance, power, and ultimately, the conclusions made, and is certainly a measurement concern (Smith, McCarthy, & Anderson, 2000). Again, it appears good practices for score reliability reporting have yet to be embraced by all researchers in applied fields.

Mean Reliability and Heterogeneity

While one inherent outcome of the RG was an assessment of certain score reliability reporting and practices in the literature for Ryff's PWB, a more specific objective was to provide a characterization of the average score reliability of the PWB. The mean coefficient alpha for the composite PWB scores was higher relative to the mean score reliabilities for the individual subscales. Although lower mean subscale alphas indicated that scores were less consistent, and more of the variance on the subscales was due to random measurement error, this finding was not surprising, as alpha is affected by the number of items, item intercorrelations, and dimensionality (Cortina, 1993). Consequently, for composite score reliability, the coefficient alpha will be an inflated value in comparison to the alphas estimated for the individual subscales (Helms et al., 2006).

Significant variability across studies was present in the coefficient alpha values for the composite, as well as all the PWB subscale scores. True heterogeneity in the mean coefficient alpha estimates across studies was confirmed with the I^2 statistic for both composite PWB and all of the subscales. This heterogeneity index confirmed that the observed variation in alphas was due to true differences, not simply a result of sampling error (Borenstein, Hedges, Higgins, & Rothstein, 2009). This heterogeneity in the coefficient alpha estimates using Ryff's PWB, reinforces the fact that reliability does not remain constant across studies. Indeed, score reliability is a dynamic property influenced by sample and test characteristics worthy of meta-analytic investigation (Vacha-Haase, 1998).

With regards to the actual values of the mean coefficient alphas for the PWB composite and respective subscales, what deems them as acceptable or adequate is dependent on the nature of the investigation, and what decisions will be made based on score interpretations (Cortina, 1993; Helms et al., 2006; Kane, 2011). As such, the author refrained from using language such as "very good" or comparing the mean score reliability values to certain thresholds for acceptable reliability. Although this is contrary to what authors of previous RG investigations have done (López-Pina et al., 2015; Schipke & Freund, 2012), score reliability is context specific and putting any sort of evaluation on coefficient alpha values may misguide and perpetuate misunderstanding. As well, many measurement experts (Cortina, 1993; Helms et al., 2006; Schmitt, 1996) have cautioned against presuming a high level of alpha ensures a measure's integrity. As Cortina (1993) aptly pointed out, just because a test may measure something *consistently*, does not give an indication of *what* is actually being measured (construct validity).

Moderator Analyses

Specific characteristics of both the sample and test can influence the score reliability (Vacha-Haase, 1998). As reliability is a function of variance, greater score variance often leads to greater score reliability (Thompson, 2003). For this reason, Thompson (2003) argued that a diverse sample may produce test scores with higher reliability. This is because more heterogeneous samples often lead to more variable scores, and thus to higher reliability. As such, moderator analyses were performed for the PWB composite and subscales to determine the specific sample and test characteristics that produced significant heterogeneity of the alphas across the studies. Specifically, age category, gender, number of items, response format, and language of the PWB were investigated as potential factors that influenced score reliability. While other potential moderators were coded for (e.g., participant health, ethnicity), the relative absence of such data from studies limited the ability to characterize the potential moderating effect of these sample variables on coefficient alpha across studies.

Age category. Significant differences in the coefficient alpha estimates were found across age categories for the composite PWB, Autonomy, Self-Acceptance, and Purpose in Life subscales. Although significant, age did not account for much of this variability as suggested by the R^2 values (ranging from 2% - 26%). A few comments may offer insight as to why age category influenced some, but not all, subscale score reliabilities. Firstly, for the purpose of moderator analysis, the average age of each study sample was used to classify each sample into a respective age category. The implication of this is that although samples participant ages may have ranged across several age categories (e.g. Diehl & Hay, 2011; Nath & Pradhan, 2012) the entire sample was categorized based on the average age. Recall that score reliability is dictated by the amount of score variance; a more heterogeneous sample would likely have greater score variance and therefore score reliability (Thompson, 2003). Additionally, marked age differences in the six aspects of the PWB were identified (Ryff, 1989; 2014), and according to the *Standards* (AERA, APA, & NCME, 2014): "when a test *is* intended to discriminate within age … populations, reliability … coefficients … should be reported separately for each subgroup" (p. 45). The fact that no consistent relationship between age category and score reliability variance was apparent across subscales, may have been a result of the difference in heterogeneity of the study samples with regards to age, and the provision of only one coefficient alpha instead of by age groups as per the *Standards* (AERA, APA, & NCME, 2014). In the RG literature, significant variability in coefficient alpha has been associated with age, with both positive and negative influences on score reliability dependent on the instrument (López-Pina et al., 2015; Therrien & Hunsley, 2013). Interestingly, in such instances, the standard deviation of age was coded and analysed as a moderator, as opposed to using age categories grouped by the average age of participants.

Gender. Gender was not found to be a moderating factor of reliability variance for either the composite PWB nor subscale score reliabilities. This is in contrast to previous RG investigations identifying gender as a significant predictor of score reliability variance (López-Pina et al., 2015; Schipke & Freund, 2012; Vacha-Haase & Thompson, 2011). The discrepancy between the current results and previous RG investigations may be due to differences in how gender was coded. For instance, López-Pina et al. (2015) coded gender as a distribution within each sample (i.e., percent male), whereas the current investigation identified studies as either all male, all female, or both regardless of the ratio of males to females. Additionally, there were few single gender, particularly, all male study samples, and instead, the majority of samples were of mixed gender. Despite these considerations, score reliability estimates for Ryff's scale do not appear to differ with respect to gender and may therefore be a sample characteristic that has trivial impact on selection of the PWB.

Number of items. As hypothesized, the number of items significantly influenced score reliability variance for the composite PWB and subscales across studies. Specifically, as the number of items increased, so too did coefficient alpha. In fact, the number of items accounted for a large proportion of the alpha variance across subscales, ranging from 40% on the Environmental Mastery subscale to 71% on the Self-Acceptance subscales. This is consistent with psychometric theory related to test length and coefficient alpha (Cortina, 1993) and reinforces its significance on score reliability as predicted by coefficient alpha. Test length was also a noteworthy predictor for score reliability in 31.2% of the RG studies review by Vacha-Haase and Thompson (2011). Similarly, Therrien and Hunsley (2013) found test length to be positively associated with reliability, with longer tests producing higher reliability estimates across various instruments measuring anxiety in older adults. Ryff (1989) herself recognized the need for balance in reducing responder burden while maintaining psychometric integrity with regards to score validity and reliability. Despite support for both the 14-item and 7-item subscale versions (Ryff, 2014), use of the 3-item version is strongly advised against, with recognized shortcomings related to reliability and validity (Ryff's, 2014). Ryff's (2014) sentiments regarding the 3-item version were re-affirmed in the current analyses, with average coefficient alphas ranging from 0.418 to 0.628 on the subscales for this extreme short form. Certainly, when selecting the PWB test version for use, researchers should consider the implications on measurement integrity. Of course, researchers must also consider practical issues related to time constraints, administrative costs, and participant burden in selecting a version format (Kruyen, Emons, & Sijtsma, 2012). Such issues are perhaps especially pertinent when conducting population based or multi-variable studies. While Smith, McCarthy, and Anderson (2000) acknowledged that in selecting a short form researchers are essentially "succumbing to temptation" (p. 102) to measure a construct in a more efficient manner, they believe that certain 'methodological sins' must still be avoided to ensure measurement rigour and responsible and ethical test use. Ultimately, researchers must recognize the potential trade-offs of various test lengths and how this will impact the interpretations that can be made.

Response format. The number of response options for Ryff's PWB significantly influenced coefficient alpha variation for composite PWB and all PWB subscale scores with the exception of the Personal Growth subscale scores. Although significant, response format did not account for any of the coefficient alpha variability as suggested by the R^2 values. One reason for the significant heterogeneity across coefficient alphas despite response formats not accounting for any of the actual variance according to the meta-regression analyses, may perhaps be a consequence of the vast majority of studies using the 6-point format.

Liu, Wu, and Zumbo (2010) suggested that the less precise a categorization of an underlying continuous variable into an ordinal scale, such as a Likert-type scale, the greater the amount of measurement error. In fact, to maximize precision and minimize bias during scale construction, Streiner and Norman (2003) outlined several factors to be considered when selecting the number of response options on a scale. Accordingly, one important consideration is the cognitive requirement that is placed on the rater to translate their answer on a particular item into one of the response options. If the number of response options does not match the raters level of discernment, there is some loss of information, and consequently, responder bias is introduced.

Despite this, the exact relationship between the number of response options and coefficient alpha remains somewhat ambiguous in the literature, with some researchers suggesting that the number of response options has no influence while others concluding that it does (Liu, Wu, & Zumbo, 2010; Weng, 2004). Regardless of the specific effect of response options on alpha, it is critical to highlight that the primary studies in the current RG, and included in the moderator analysis, used response formats ranging from 4-point to 7-point categories, in contrast to the 6-point original (Ryff, 1989). During scale construction, Ryff (1989) evaluated items based on a number of criteria including the ability of each item to produce variable responses on the 6-point scale ranging from strongly agree to strongly disagree. Likely, the 6-point scale was selected purposefully to address some of the aforementioned discernment issues. As significant differences in mean coefficient alphas across the various number of response categories were indeed observed in the current study, test users may want to take this into account prior to modifying the response format from the originally developed version.

Language of test. As hypothesized, language significantly influenced coefficient alpha variation for composite PWB and all PWB subscale scores. Schipke and Freund (2012) similarly found that language significantly affected coefficient alphas for the Physical Self-Description Questionnaire (PSDQ) scores in their RG meta-analyses, although they reported consistently lower score reliabilities with translated PSDQ versions. Schipke and Freund (2012) dichotomously coded this variable, so that all versions besides the original English version were coded as “translated”, which may account for the somewhat different results in the current investigation, in which each language version was separately analyzed.

The fact that language may impact score reliability on the PWB, is in keeping with Kline's (2009) comment that modifying a test in any way, including the adaption of a test into another language, changes the conditions of measurement, which may influence the score reliability. Ironically, the appeal of translating an already existing test is the presumption that it "carries with it the scientific integrity and potentially rich theoretical and psychometric history of the original test" (Zumbo, 2003, p. 136). Yet, researchers must ensure measurement equivalence, with respect to item content as well as score validity and reliability across the test versions (Byrne et al., 2009; Zumbo, 2003).

With regards to the findings in the current investigation, several language adaptions maintained comparable score reliability estimates with the original English version (i.e. Italian translation), while others reported very low coefficient alphas (i.e. Swedish translation). Upon inspection of the Italian and Swedish translations, however, it became apparent that the Italian translation of the PWB adopted 14 items per subscale, while the Swedish version was shorter in length with only 3 items per subscale. As a result of the type of moderator analyses run in the present investigation, it was not possible to determine whether it was the actual language version or number of items that influenced the score reliabilities, as test characteristics could only be investigated individually. Yet comparing the R^2 values of the meta-regressions, language explained less of the variance in the coefficient alphas in comparison to test length and appears to be less meaningful than the number of items on the PWB with respect to influencing coefficient alpha.

For cross-cultural test adaptation, measurement experts have most recently advocated for the use of item response theory (IRT) over the classical test model, as the later item statistics, including coefficient alpha, are dependent on sample characteristics and more prone to fluctuate across cultures (Byrne et al., 2009; Zumbo, 2003). Additionally, based on the current findings, different translations of Ryff's PWB use various numbers of items, and of course, the length of the test also affects coefficient alpha (Cortina, 1993). Consequently, Hambleton, Bartram, and Oakland (2011) suggested that the modern measurement framework, IRT, and associated methods are more appropriate for test construction, adaption, and evaluation, including the estimation of score reliability.

Limitations

It is critical to highlight the limitations of the current investigation. Firstly, reliability can arguably be conceptualized two ways (Knapp & Sawilowsky, 2001; Sawilowsky 2000a; 2000b). As suggested by the editorial policies put forth by *Educational and Psychological Measurement*, Thompson (1994), Vacha-Haase (1998), and Thompson and Vacha-Haase (2000), the notion that reliability should be conceptualized as a property of test *scores* represents the datacentric view. In taking this perspective, there are problems with inconsistent terminology in the measurement literature, as well as in test selection (Sawilowsky 2000a; 2000b). Instead, Sawilowsky (2000a; 2000b) contended that the psychometric concept of reliability refers to the instrument, and as such, researchers should cite the reliability estimate from the test manual in addition to the reliability estimates of their own sample's scores complete with descriptive statistics. While the author of this investigation does not assert measurement expertise, she acknowledges the merits inherent in each perspective, and attempted to incorporate both into the current RG investigation.

Secondly, Sijtsma and van der Ark (2015) aptly point to the fact that CTT is but one model for conceptualizing reliability, only accounting for one source of error at a time (Dimitrov, 2002). For the purposes of the current investigation, the focus was on reliability as defined according to CTT as the reliability coefficient most commonly estimated and reported with regards to Ryff's scale, is coefficient alpha, which lies within the CTT framework. Consequently, this narrowed scope for the current investigation, did not address all sources and estimates of measurement error that would be worthy to examine with regards to Ryff's scale.

In selecting coefficient alpha (Cronbach, 1951) as the reported reliability estimate to investigate, the limitations of this particular statistic must also be highlighted. As previously mentioned, coefficient alpha is a realistic estimate of reliability only if certain assumptions are met regarding true and error scores (Dimitrov, 2002; Yang & Green, 2015). Specifically, if either assumption of essential tau-equivalency or uncorrelated errors is violated, the consequence is a biased reliability estimate (Yang & Green, 2015). In fact, Sijstma (2009a; 2009b) argued that coefficient alpha is only equal to score reliability under impractical conditions, and is instead, a lower bound to the score reliability. As such, Sijstma (2009a; 2009b) maintained that there are more useful and realistic score reliability estimates. Whether the primary studies utilized in this RG meta-analysis acknowledged and tested for these assumptions related to coefficient alpha were unknown as none of the researchers explicitly reported this, and consequently, the alpha values generalized across studies in this investigation, may have been biased by such factors.

Problems with RG analysis may also result from using alpha coefficients reported in primary studies that have used different versions of the instrument (Dimitrov, 2002; Rodriguez & Maeda, 2006). Various lengths, response formats, scale intervals, contexts of age, language, and culture, in addition to the sample (random versus nonrandom) will influence coefficient alpha, and will undoubtedly impact the generalization of alpha across the studies (Dimitrov, 2002; Rodriguez & Maeda, 2006). Yet, if primary studies fail to report this information, the RG researcher is unable to address such issues (Dimitrov, 2002; Henson & Thompson, 2002; Thompson and Vacha-Haase, 2000). As Thompson and Vacha-Haase (2000) maintained "the RG chef can only work with the ingredients provided by the literature" (p. 184). As such, the author erred on the side of caution, and unless authors explicitly reported information regarding sample and test characteristics, the variable remained blank for that particular study and consequently, was not included in the RG analyses.

Researchers may want to consider how data is retrieved, coded, and analyzed, as these factors will influence the study outcome (Sánchez-Meca et al., 2013; Ioannidis, 2010). As unpublished studies were excluded from the meta-analysis in the current study, this RG analysis may be biased (Howell & Shields, 2008). This investigation was also limited by the data collection period in addition to the explicit keywords utilized in the search strategy. Consequently, these findings may not be representative of all RG studies for Ryff's scale if alternate dates and search parameters are used. With regards to coding, sample age was recorded as average age, and then converted to into age categories for moderator analyses. Nuances in sample participant age across studies was lost by using the mean age to categorize samples. Coding related to gender and language additionally influenced how data was analyzed and prevented comparison to previous RG work that used different coding methods.

Future Recommendations

Despite the aforementioned limitations, interpretation of these findings leads to a number of recommendations that are worth highlighting.

Competencies of the test user and considerations for responsible test use. The *Standards for Educational and Psychological Testing* dedicated an entire chapter to recommendations related to the rights and responsibilities of test users **(**AERA, APA, & NCME, 1999, 2014). Of course, the assumption is that test users have the underlying competence, including the knowledge and training, to use tests appropriately, ethically and professionally (AERA, APA, & NCME, 1999, 2014; International Test Commission, 2001). Results of the current investigation highlight the importance of test user knowledge in basic psychometric principles and procedures related to score reliability. In fact, an instrument’s conceptual and structural properties, length, language, and response options, as well as sample characteristics, may affect score reliability, particularly coefficient alpha, and researchers should take such issues into consideration when selecting the particular test version for use. Researchers must recognize the potential influence of test and sample characteristics on score reliability estimates, and how this may impact the interpretations that can be made. Despite the foundational importance of measurement for sound research practice it is receiving less coverage in graduate programs today (Kline, 2009; Reynolds, 2010; Thompson, 2003).

Score reliability and test adaption reporting. Although this study used a particular instrument (Ryff's PWB) to investigate the treatment of reliability in applied research, it reinforces the fact that there is still ample room to improve score reliability reporting practices in the literature. Specifically, authors of primary articles should *explicitly* report details related to the sample characteristics used in their research with which they calculate score reliability. Similarly, if scores are reported for more than one group or time point, and these scores will be used in the analyses and interpreted, score reliability should be reported for each score. The specifics of the type of reliability estimate should also be detailed (e.g. coefficient alpha). As score reliability is also influenced by the specific test characteristics, the reporting of such test factors (e.g. number of items, response format, etc.) are also important. Finally, when a test is altered in any way, documentation of these changes should be presented (the *Standards* (AERA, APA, & NCME, 1999, 2014); the APA Task Force (Wilkinson & APA Task Force on Statistical Inference, 1999)). In fact, even if no language translation is necessary, when tests developed for use in one culture are used in another cultural context, modifying a test may be necessary in which specific adaptation guidelines and reporting of modifications apply (Hambleton, Bartram, & Oakland (2011). Adherence to such guidelines is expected for responsible test use (AERA, APA, & NCME, 1999, 2014; Byrne et al., 2009). Of course, an additional benefit of such detailed reporting in primary reports is that this provides more information for RG investigations (Vacha-Haase & Thompson, 2011).

Conclusions

This investigation highlighted the dynamic nature of score reliability and the current reporting practices in peer-reviewed journals using the specific example of Ryff's Scale of Psychological Well-Being. In keeping with measurement experts (Hambleton, Bartram, & Oakland, 2011; Kline, 2009; Vacha-Haase & Thompson, 2011) and current guidelines (the *Standards*, APA Task Force on Statistical Inference), there are necessary duties and responsibilities for test users, including explicit reporting of the instrument itself and score reliability. Not only is this critical for score validity and the inferences made, but returning to Nunally's remarks nearly thirty-five years ago, research is restricted "by the reliability of measuring instruments and by the reliability with which scientists use them" (p. 1589, Nunally, 1982). Consequently, improving the understanding of reliability and the minimum requirements for test and score reliability reporting should be of paramount concern to all those involved in the research process. Vacha-Haase's (1998) proposed Reliability Generalization directly addresses score reliability issues, and in effect, attention to measurement integrity and research rigor.

The findings in the current investigation indicate significant variability in mean score reliability for Ryff's PWB composite as well as subscales, and reiterates the importance of researchers reporting score reliability for their own sample's data. Moderator analyses indicated that significant differences in mean coefficient alphas were most apparent for different PWB versions with regards to test length, although other features of the instrument appear to influence score reliability as well. This study, and other such applied RG studies, will further score reliability discourse, by providing considerations for test use and reporting practices, as well as confronting reliability myths and misunderstanding that continue to persist.

References

Abbott, R. A., Ploubidis, G. B., Huppert, F. A., Kuh, D., Wadsworth, M. E. J., & Croudace, T. J. (2006). Psychometric evaluation and predictive validity of Ryff's psychological well-being items in a UK birth cohort sample of women. *Health and Quality of Life Outcomes*, *76*. doi:10.1186/1477-7525-4-76

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Archana, Updesh, K., & Singh, R. (2014). Resilience and spirituality as predictors of psychological well-being among university students. *Journal of Psychosocial Research*, *9*, 227-235.

Aschbrenner, K. A., Greenberg, J. S., Seltzer, M. M. (2009). Parenting an adult child with bipolar disorder in later life. *Journal of Nervous and Mental Disease*, *197*, 298-304. doi: 10.1097/NMD.0b013e3181a206cc

Awan, S., & Sitwat, A. (2014). Workpace spirituality self-esteem and psychological well-being among mental health professionals. *Pakistan Journal of Psychological Research*, *29*, 1-12.

Barry, A. E., Chaney, B. H., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behaviour: A review of seven journals. *Health Education & Behaviour, 41*, 12-18. doi: 10.1177/1090198113483139

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, *62*, 254-263. doi: 10.1177/0013164402062002004

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*, 335–340. doi:10.3102/10769986027004335

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009) *Introduction to meta-analysis*. West Sussex, United Kingdom: John Wiley & Sons Limited.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive Meta-Analysis (Version 3.0) [computer software]. New Jersey: Biostat.

Boylan, J. M., & Ryff, C. D. (2015). Psychological well-being and metabolic syndrome: Findings from the midlife in the United States National Sample. *Psychosomatic Medicine, 77*, 548-558. doi: 10.1097/PSY.0000000000000192

Byrne, B. M., Oakland, T., Leong, F. T. L., van de Vijver, F. J. R.; Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3*, 94-105. doi: 10.1037/a0014516

Caruso, J. (2000). Reliability generalization of the Neo Personality Scales. *Educational and Psychological Measurement 60*, 236-254. doi: 10.1177/00131640021970484

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*, 287-292.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98-104. Retrieved from http://eds.a.ebscohost.com/

Clarke, P. J., Marshall, V. W., Ryff, C. D., & Wheaton, B. (2001). Measuring psychological well-being in the Canadian Study of health and aging. *International Psychogeriatrics*, *13*, 79-90. doi: 10.1017/S1041610202008013

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.

Diehl, M., & Hay, E. (2011). Self-concept differentiation and self-concept clarity across adulthood: associations with age and psychological well-being. *International Journal of Aging and Human Development*, *73*, 125-152. doi: 10.2190/AG.73.2.b

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement*, *62*, 783-801. doi: 10.1177/001316402236878

Franz, C. E., Panizzon, M. S., Eaves, L. J., Thompson, W., Lyons, M. J., Jacobson, K. C., . . . Kremen, W. S. (2012). Genetic and environmental multidimensionality of well- and ill-being in middle aged twin men. *Behaviour Genetics*, *42*, 579-591. doi:

10.1007/s10519-012-9538-x

Fava, G. A., Rafanelli, C., Ottolini, F., Ruini, C., Cazzaro, M., & Grandi, S. (2001). Psychological well-being and residual symptoms in remitted patients with panic disorder andagoraphobia. *Journal of Affective Disorders, 65*, 185-190. doi: 10.1016/S0165-0327(00)00267-6

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment*, *30*, 130-139. doi: 10.1027/1015-5759/a000181

Green, C. E., Chen, C. E., Helms, J. E., & Henze, K. T. (2011). Recent reliability reporting practices in *Psychological Assessment*: Recognizing the people behind the data. *Psychological Assessment*, *23*, 656-669. doi: 10.1037/a0023089

Green, S. B, & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice*, *34*, 14–20. doi: 10.1111/emip.12100

Hamilton, R. K., Bartram, D., Oakland, T. (2011). Technical advances and guidelines for improving testing practices. In P. R. Martin, F. M. Cheung, M. C. Knowles, M. Kyrios, L. Littlefield, J. B. Overmier, and J. M. Prieto. (Eds.), *IAAP Handbook of Applied Psychology* (pp. 338 – 361). Blackwell Publishing Limited.

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika*, *41*, 219–231. doi:10.1007/BF02291840

Henson, R. K. Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189. Retrieved from http://eds.a.ebscohost.com/

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "Reliability Generalization" studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-126. Retrieved from http://eds.a.ebscohost.com/

Helms, J. E., Henze, K. T., Sass, T. L., & Mifsud, V. A. (2006). Treating Cronbach's alpha reliability coefficients as data in counselling research. *The Counselling Psychologist*, *34*, 630-660. doi: 10.1177/0011000006288308

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*, 523-531. doi: 10.1177/00131640021970691

Howell, R. T., & Shields, A. L. (2008). The file drawer problem in reliability generalization: A strategy to compute a Fail-Safe N with reliability coefficients. *Educational and Psychological Measurement*, *68*, 120-128. doi: 10.1177/0013164407301528

Huedo-Medina, T. B., Sánchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or I^2? *Psychological Methods*, *11*, 193-206. doi: 10.1037/1082-989X.11.2.193

Huta, V., & Waterman, A. S. (2014). Eudaimonia and its distinction from hedonia: Developing a classification and terminology for understanding conceptual and operational definitions. *Journal of Happiness Studies*, *15*, 1425-1456. doi: 10.1007/s10902-013-9485-0

International Test Commission (2001). International guidelines for test use. *International Journal of Testing*, *1,* 93-114. Retrieved from https://www.intestcom.org

Ioannidis, J. P. A., (2010). Meta-research: The art of getting it wrong. *Research Synthesis Methods*, *1,* 169-184. doi: 10.1002/jrsm.19

Joshanloo, M., & Ghaedi, G. (2009). Value priorities as predictors of hedonic and eudaimonic aspects of well-being. *Personality and Individual Differences*, *47,* 294-298. doi: 10.1016/j.paid.2009.03.016

Kafka G. J., Kozma, A. (2002). The construct validity of Ryff's Scales of Psychological Well-Being (SPWB) and their relationship to measures of subjective well-being. *Social Indicators Research*, *57*, 171–190. Retrieved from http://journals1.scholarsportal.info/

Kaur, S., & Gupta, S. (2013). Effect of religiosity on psychological well-being of senior citizens living with family and living in old age homes. *Journal of Psychosocial Research*, *8*, 209-222.

Kline, R. B. (2009). Measurement. *Becoming a behavioural science researcher: A guide to producing research that matters* (pp. 191-222). New York, NY: The Guilford Press.

Knapp, T. R. (1977). The reliability of a dichotomous test-item: A "Correlationless" Approach. *Journal of Educational Measurement*, *14,* 237-252. Retrieved from http://www.jstor.org/

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*, 2693-2710. Doi: 10.1002/sim.1482

Knapp, T. R., & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education*, *70*, 65-79. Retrieved from http://www.jstor.org/

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, *12*, 321-344. doi: 10.1080/15305058.2011.643517

Leue, A., & Lange, S. (2011). Reliability generalization: An examination of the positive affect and negative affect schedule. *Assessment*, *18*, 487-501. doi: 10.1177/1073191110374917

Liu, Y., Wu, A. D., & Zumbo, B. D. (2010). The impact of outliers on Cronbach's coefficient alpha estimate of reliability: Ordinal/rating scale item responses. *Educational and Psychological Measurement*, *70*, 5-21. doi: 10.1177/0013164409344548

López-López, J. A., Botella, J., Sánchez-Meca, J., & Marín-Martínez, F. (2013). Alternatives for mixed-effects meta-regression models in the reliability generalization approach: A simulation study. *Journal of Educational and Behavioral Statistics*, *38*, 443-469. doi: 10.3102/1076998612466142

López-Pina, J. A., Sánchez-Meca, J., López-López, J. A., Marín-Martínez, F., Núñez-Núñez, R. M., Rosa-Alcazar, A. I., Gómez-Conesa, A., & Ferrer-Requena, J. (2015). The Yale-Brown Obsessive Compulsive Scale: A reliability generalization meta-analysis. *Assessment*, *22*, 619-628. doi: 10.1177/1073191114551954

Mack, D. E., Wilson, P. M., Gunnell, K. E., Gilchrist, J. D., Kowalski, K. C., & Crocker, P. R. E. (2012). Health-enhancing physical activity: Associations with markers of well-being. *Applied Psychology: Health and Well-Being, 4,* 127-150. doi: doi:10.111/j.1758-0854.2012.01065.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749. doi: 10.1037/0003-066X.50.9.741

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., . . . PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *British Medical Journal*, 349. doi: 10.1136/bmj.g7647

Nath, P., & Pradhan, R. K. (2012). Influence of positive affect on physical health and psychological well-being: Examining the mediating role of psychological resilience. *Journal of Health Management*, *14*, 161-174. doi: 10.1177/097206341201400206

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw Hill.

Nunnally, J. C. (1982). Reliability of measurement. In H. E. Mitzel (Ed.), *Encyclopedia of educational research* (pp. 1589-1601). New York, NY: Free Press.

Rexrode, K. R., Petersen, S., & O'Toole, S. (2008). The Ways of Coping Scale: A Reliability Generalization study. *Educational and Psychological Measurement*, *68*, 262-280. doi: 10.1177/0013164407310128

Reynolds, C. R. (2010). Measurement and assessment: An editorial view. *Psychological Assessment*, *22*, 1-4. doi: 10.1037/a0018811

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*, 306-322. doi: 10.1037/1082-989X.11.3.306

Ryan, M. R., & Deci, E. L. (2001). On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology*, *52,* 141-166. Retrieved from http://eds.a.ebscohost.com/

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, *57*, 1069-1081. doi: 10.1037/0022-3514.57.6.1069

Ryff, C. D. (2013). Eudaimonic well-being and health: Mapping consequences of self-realization. In A. S. Waterman (Ed.), *The best within us: Positive psychology perspectives on eudaimonia* (pp. 77-98). Washington, DC: American Psychological Association.

Ryff, C. D. (2014). Psychological well-being revisited: Advances in the science and practice of eudaimonia. *Psychotherapy and Psychosomatics, 83*, 10-28. doi: 10.1159/000353263

Ryff, C. D., & Singer, B. H. (2008). Know thyself and become what you are: A eudaimonic approach to psychological well-being. *Journal of Happiness Studies, 9*, 13-39. doi: 10.1007/s10902-006-9019-0

Sánchez-Meca, J., López-López, J. A., & López-Pina, J.A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology*, *66*, 402-425. doi:10.1111/j.2044-8317.2012.02057.x

Sawilowsky, S.S. (2000a). Psychometrics versus datametrics: Comment on Vacha-Haase's "Reliability Generalization" method and some *EPM* editorial policies.

Educational and Psychological Measurement, *60*, 157-173. doi: 10.1177/00131640021970439

Sawilowsky, S.S. (2000b). Reliability: Rejoinder to Thompson and Vacha-Haase. *Educational and Psychological Measurement*, *60*, 196-200. doi: 10.1177/00131640021970457

Schipke, D., & Freund, P. A. (2012). A meta-analytic reliability generalization of the Physical Self-Description Questionnaire (PSDQ). *Psychology of Sport and Exercise*, *13*, 789-797. doi: 10.1016/j.psychsport.2012.04.012

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350-353. doi: 10.1037/1040-3590.8.4.350

Schultz, K. S., & Whitney, D. J. (2005). Module 5: Classical true score theory and reliability. *Measurement Theory in Action* (pp. 69-85). Thousand Oaks, CA: Sage.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120. doi: 10.1007/s11336-008-9101-0

Sijtsma, K. (2009b). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing*, *9*, 167-194. doi: 10.1080/15305050903106883

Sijtsma, K., & Emons, W. H. M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, *70*, 565- 572. doi: 10.1016/j.jpsychores.2010.11.002

Sijtsma, K., & van der Ark, L. A. (2015). Conceptions of reliability revisited and practical recommendations. *Nursing Research*, *64*, 128-136. doi: doi.org.proxy.library.brocku.ca/10.1097/NNR.0000000000000077

Siconolfi, D. E., Halkitis, P. N., Barton, S. C., Kingdon, M. J., Perez-Figueroa, R. E., Arias-Martinez, V., … Brennan-Ing, M. (2013). Psychosocial and demographic correlates of drug use in a sample of HIV-positive adults ages 50 and older. *Prevention Science*, *14*, 618-627. doi: 10.1007?s11121-012-0338-6

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*, 102-111. doi: 10:1037//1040-3590.12.1.102

Springer K. W., Hauser R. M, & Freese, J. (2006). Bad news indeed for Ryff's six-factor model of well-being. *Social Science Research*, *35*,1120–1131. doi: doi:10.1016/j.ssresearch.2006.01.003

Springer, K. W., Pudrovska, T., Hauser, R. M. (2011). Does psychological well-being change with age? Longitudinal tests of age variations and further exploration of the multidimensionality of Ryff's model of psychological well-being. *Social Science Research*, *40*, 392-398. doi:10.1016/j.ssresearch.2010.05.008

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677-680. Retrieved from http://www.jstor.org/

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*, 99-103. Retrieved from http://journals1.scholarsportal.info/

Therrien, Z., & Hunsley, J. (2013). Assessment of anxiety in older adults: A reliability generalization meta-analysis of commonly used measures. *Clinical Gerontologist*, *36*, 171-194. doi: 10.1080/07317115.2013.767871

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development*, *70*, 434-438. Retrieved from eds.b.ebscohost.com

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54,* 837-847.

Thompson, B. (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195. doi: 10.1177/00131640021970439

Thye, S. R. (2000). Reliability in experimental sociology. *Social Forces*, *78*, 1277-1309.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*, 8-14.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20. doi: 10.1177/0013164498058001002

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*, 562-569. doi: 10.1177/0013164402062004002

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those test manuals: Validity of scores reliability inductions. *Educational and Psychological Measurement, 60,* 509-522. doi: 10.1177/00131640021970682

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, *44,* 159-168. doi: 10.1177/0748175611409845

Vescovelli, F., Albieri, E., & Ruini, C. (2014). Self-rated and observer-rated measures of well-being and distress in adolescence: an exploratory study. *SpringerPlus*, *3*, 490. doi: 10.1186/2193-1801-3-490

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*, 956-972. doi: 10.1177/0013164404268674

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Wilson, P. M., Mack, D. E., & Sylvester, B. D. (2011). When a little myth goes a long way: The use (or misuse) of cut-points, interpretations, and discourse with coefficient-alpha in exercise psychology. In A. M. Columbus (ed.), *Advances in Psychology Research* (pp. 1-17). Nova Science Publishers.

Yang, Y., & Green, S. B. (2015). Further discussion on reliability: The art of reliability estimation. *Nursing Research*, *64*, 146-151. doi: 10.1097/NNR.0000000000000080

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability Generalization across studies. *Educational and Psychological Measurement*, *60*, 201-223. doi: 10.1177/00131640021970466

Zumbo, B. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*, 136–147. doi: 10.1191/0265532203lt248oa

Appendix A

Factors Affecting Score Reliability

<u>Characteristics of Test</u>

- Test length (test items)
- Range of item difficulty
- Evenness in scaling
- Item dependence
- Scoring objectivity
- Scoring inaccuracy
- Chance in getting correct answer
- Position of the correct item among alternatives
- Homogeneity of test content
- Item wording
- Item length
- Item clarity
- Test instructions

<u>Characteristics of Test-taker</u>

- Test content familiarity
- Test taking speed
- Test taking accuracy
- Test incentive
- Individual effort
- Illness/Affect
- Cheating

<u>Circumstantial Characteristics</u>

- Test distractions
- Accidents during test examination
- Testing period in relation to calendar date

Appendix B

Definitions of Theory-Guided Dimensions of Well-Being

Self-acceptance High scorer: Possesses a positive attitude toward the self; acknowledges and accepts multiple aspects of self including: good and bad qualities; feels positive about past life. Low scorer: Feels dissatisfied with self; is disappointed with what has occurred in past life; is troubled about certain personal qualities; wishes to be different than what he or she is.
Positive relations with others High scorer: Has warm, satisfying, trusting relationships with others; is concerned about the welfare of others, capable of strong empathy, affection, and intimacy; understands give and take of human relationships. Low scorer: Has few close, trusting relationships with others; finds it difficult to be warm, open, and concerned about others; is isolated and frustrated in interpersonal relationships; not willing to make compromises to sustain important ties with others.
Autonomy High scorer: Is self-determining and independent; able to resist social pressures to think and act in certain ways; regulates behaviour from within; evaluates self by personal standards. Low scorer: Is concerned about the expectations and evaluations of others; relies on judgments of others to make important decisions; conforms to social pressures to think and act in certain ways.
Environmental mastery High scorer: Has a sense of mastery and competence in managing the environment; controls complex array of external activities; makes effective use of surrounding opportunities; able to choose or create contexts suitable to personal needs and values. Low scorer: Has difficulty managing everyday affairs; feels unable to change or improve surrounding context; is unaware of surrounding opportunities; lacks sense of control over external world.
Purpose in life High scorer: Has goals in life and a sense of directedness; feels there is meaning to present and past life; holds beliefs that give life purpose; has aims and objectives for living. Low scorer: Lacks a sense of meaning in life; has few goals or aims, lacks sense of direction; does not see purpose of past life; has no outlook or beliefs that give life meaning.
Personal growth High scorer: Has a feeling of continued development; sees self as growing and expanding; is open to new experiences; has sense of realizing his or her potential; sees improvement in self and behaviour over time; is changing in ways that reflect more self-knowledge and effectiveness.

Low scorer: Has a sense of personal stagnation; lacks sense of improvement or expansion over time; feels bored and uninterested with life; feels unable to develop new attitudes or behaviours.

(Ryff, 1989, p. 1072)

Appendix C

Search Strategy Utilized in PsycINFO Database

The following keywords were entered into PsycINFO: ((eudaimonic OR eudaemonic) AND “well-being”) OR ryff OR “scale* of psychological well-being” OR “psychological well-being scale*” OR “model of psychological well-being”. *Any Field* was selected as the parameters for locating the keywords. The search was further limited to peer reviewed journals. Under the *Tests and Measures,* “Psychological Well-Being Scale” was selected, as PsycINFO allows for further refinement this way. Publication date was set from 1989 onwards. Last PsycINFO search was February 18th, 2016.

Appendix D

RG Coding Form

STUDY IDENTIFIERS

Column ___: PUB_YEAR:
(Numeric year article was published or written if dissertation)

Column ___: AUTHOR(S):
(Last name only)

SAMPLE CHARACTERISTICS

Column ___: SAMPLE_SIZE

Column ___: AGE (average)

Did not report

Column ___: AGE CATEGORY

Defined as:
CHILDREN = 0 to 12 years
ADOLESCENTS = 13 to 18 years
EMERGING ADULTS = 19 to 25 years
ADULTS = 25 to 65 years
OLDER ADULTS = 65 years and older

Column ___: GENDER

(1) Male

(2) Female
(3) Mixed

Column ___: SAMPLE_ETHNICITY

Column ___: SAMPLE_HEALTH

TEST CHARACTERISTICS

Column ___: NUMBER_OF_ITEMS_PER_SUBSCALE

Column ___: RESPONSE_FORMAT

Column ___: LANGUAGE_OF_PWB

STUDY RELIABILITY

Column ___: COEFFICIENT_ALPHA_SELF-ACCEPTANCE

Column ___: COEFFICIENT_ALPHA_ AUTONOMY

Column ___: COEFFICIENT_ALPHA_PERSONAL_GROWTH

Column ___: COEFFICIENT_ALPHA_PURPOSE IN LIFE

Column ___: COEFFICIENT_ALPHA_ENVIRONMENTAL MASTERY

Column ___: COEFFICIENT_ALPHA_POSTIVE RELATIONS

Table 1.
Sample and Test Characteristics for Corresponding Coefficient Alpha Estimates

Sample Characteristics	Composite PWB (*k* = 150)	AU Subscale (*k* = 113)	EM Subscale (*k* = 118)	PG Subscale (*k* = 121)	PR Subscale (*k* = 122)	PL Subscale (*k* = 137)	SA Subscale (*k* = 120)
Age Category							
Children		1	1	1	1	1	1
Adolescents	15	10	11	10	11	11	10
Emerging Adults	51	34	34	34	34	35	37
Adults	67	46	47	52	51	54	50
Older Adults	10	11	9	12	10	20	9
Did not report	7	11	16	12	15	16	13
Gender							
Female	15	8	11	9	11	12	9
Male	8	3	3	2	4	3	5
Both	125	100	102	108	105	120	104
Did not report	2	2	2	2	2	2	2
Ethnicity							
African			1		1	1	
African American	8						
Arab or Arab Canadian	2				2		
Caucasian					1	4	1
Columbian		1	2	1		1	1
European	1	1	1	1			
Filipino	1						
Korean	1						
Latino	1		1		1		
Spanish	1	1	2	2	2	1	1
Mixed	58	32	31	36	35	41	34

Did not report	77	78	80	81	80	89	83
Sample Health							
Cancer (unspecified)	1						
Breast Cancer	1						
Prostate Cancer	1						
Multiple Sclerosis	1		1			1	1
Bi-polar (I or II)		1	1	1	1	1	1
Depression	2	1	1	1	1	1	1
History of Mood Disorder		1	1	1	1	1	1
Mental Illness (unspecified)	2	1	1	1	1	1	1
Neck and Shoulder Pain	1	1	1	1	1	1	1
Physical Disability		1	1	1	1	1	1
HIV Positive		1	2		1	1	2
Burnout	2						
Hearing impairment	1						
Mixed	7	4	7	6	8	13	4
Did not report	127	94	95	100	100	104	100
Other	6	8	7	9	7	12	7
Test Characteristics for PWB							
Number of Items							
3		20	22	21	19	20	22
4			2	4	2	2	7
5		2	9	4	7	7	1
6	2	4	9	2	9	10	9
7	1	5	5	15	7	13	7
8		10				1	
9		20	17	18	18	22	18
12		1					

13					1		
14		32	31	37	37	39	35
15	1						
18	68	1	1	1	1	1	1
19	2						
20	3	1	1	1	1	2	1
22	2						
24	14						
28	1						
29	3						
30	1						
31	1						
36	2						
42	8						
46	1						
54	12						
60	1						
84	19						
86	1						
Did not report	6	17	21	18	20	20	19
Response format							
2-point		2	2	2	2	2	2
4-point	6	2	3	3	3	3	3
5-point	28	4	5	4	7	5	6
6-point	65	72	71	76	75	86	73
7-point	13	12	13	13	14	15	13
8-point	1						
Did not report	36	21	24	23	21	26	23
Language							

Chinese	7	1	1	1	1	1	2
English	78	54	53	59	59	70	55
Filipino		1	1	1	1	1	1
French				1	1	3	
German	1	1	1	1	1	1	1
Greek	1						
Iranian	5	1	1	1	1	1	1
Italian	1	6	6	6	7	6	7
Japanese	2	3	3	3	3	3	3
Korean	1	2	2	2	2	2	2
Malay	1	1	1	1	1	1	1
Mandarin	2	1	1	1	1	1	1
Persian	1						
Polish	5						
Portuguese	5	4	4	5	4	4	5
Spanish	7	14	18	16	17	15	15
Swedish	3	5	5	5	5	5	5
Turkish		1	1	2	1	2	1
Urdu		1	1	1	1	1	1
More than one language	2						
Did not report	27	17	19	15	16	20	19

Note. k = number of reliability coefficients

Table 2.
Overall Reliability and 95% Confidence Intervals for the Alpha Coefficients Across All Studies Employing PWB

Scale/Subscales	*k*	*Mean*	*Min.*	*Max.*	*95% CI [Lb, Ub]*	*Q*	I^2
Coefficient α							
PWB - Composite	150	0.858	0.51	0.99	[0.846, 0.869]	5271.57**	97.17
PWB - Autonomy	113	0.722	0.13	0.90	[0.697, 0.745]	3485.30**	96.79
PWB - Environmental Mastery	118	0.728	0.00	0.92	[0.705, 0.750]	3248.89**	96.40
PWB - Personal Growth	121	0.729	0.18	0.91	[0.706, 0.750]	3633.64**	96.70
PWB - Positive Relations	122	0.775	0.23	0.93	[0.754, 0.795]	3970.29**	96.95
PWB - Purpose in Life	137	0.750	-0.35	0.97	[0.726, 0.772]	7406.17**	98.16
PWB - Self-Acceptance	120	0.801	0.17	0.95	[0.780, 0.819]	4433.86**	97.32

Note. *k* = number of reliability coefficients; *Min* = lowest reliability coefficient; *Max.* = highest reliability coefficient; *95% CI [Lb, Ub]* = lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; *Q* = heterogeneity statistic; I^2 = heterogeneity index
** $p < .001$

Table 3.
Moderator Analyses for Alpha Coefficients Across All Studies Adopting Composite Scale

Covariate	*k*	*Mean*	*Min.*	*Max.*	*95%CI [Lb, Ub]*	*Q*	*p*	I^2
Age Category						31.43	0.000	97.27
Adolescents	15	0.810	0.65	0.85	[0.790, 0.827]			
Emerging Adults	51	0.877	0.51	0.97	[0.860, 0.892]			
Adults	67	0.848	0.68	0.97	[0.826, 0.867]			
Older Adults	10	0.908	0.66	0.99	[0.816, 0.955]			
Gender						5.15	0.076	97.20
Female	15	0.841	0.68	0.95	[0.800, 0.875]			
Male	8	0.824	0.74	0.93	[0.788, 0.855]			
Both	125	0.861	0.51	0.99	[0.848, 0.873]			
Number of Items						148.29	0.000	97.06
18	68	0.794	0.65	0.91	[0.783, 0.804]			
24	14	0.846	0.66	0.91	[0.822, 0.868]			
29	3	0.882	0.86	0.91	[0.840, 0.913]			
42	8	0.890	0.81	0.92	[0.859, 0.914]			
54	12	0.913	0.51	0.99	[0.883, 0.935]			
84	19	0.939	0.77	0.97	[0.921, 0.953]			
Response Format						8.61	0.035	96.38
4-point	6	0.825	0.68	0.97	[0.534, 0.941]			
5-point	28	0.854	0.72	0.95	[0.832, 0.873]			
6-point	65	0.854	0.51	0.97	[0.836, 0.871]			
7-point	13	0.811	0.73	0.91	[0.781, 0.837]			
Language						52.46	0.000	97.62
Chinese	7	0.871	0.77	0.94	[0.805, 0.915]			
English	78	0.867	0.51	0.99	[0.852, 0.881]			
Italian	5	0.780	0.73	0.83	[0.738, 0.815]			
Portuguese	5	0.890	0.70	0.97	[0.738, 0.956]			
Spanish	5	0.887	0.86	0.91	[0.867, 0.904]			

Swedish	7	0.768	0.69	0.83	[0.715, 0.811]
Turkish	3	0.851	0.77	0.93	[0.700, 0.929]

Note. *k* = number of reliability coefficients; *Min* = lowest reliability coefficient*; Max.* = highest reliability coefficient; *95% CI [Lb and Ub]* = lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; *Q* = heterogeneity statistic; I^2 = heterogeneity index

Table 4.
Moderator Analyses for Alpha Coefficients Across All Studies Adopting PWB Autonomy Subscale

Covariate	*k*	*Mean*	*Min.*	*Max.*	*95%CI [Lb, Ub]*	*Q*	*p*	I^2
Age Category						10.39	0.015	97.02
Adolescents	10	0.641	0.13	0.86	[0.516, 0.740]			
Emerging Adults	34	0.770	0.55	0.90	[0.735, 0.801]			
Adults	46	0.717	0.36	0.88	[0.680, 0.750]			
Older Adults	11	0.671	0.29	0.85	[0.568, 0.754]			
Gender						1.76	0.415	96.80
Female	8	0.762	0.58	0.85	[0.689, 0.740]			
Male	3	0.772	0.68	0.88	[0.580, 0.883]			
Both	100	0.715	0.13	0.90	[0.689, 0.740]			
Number of Items						117.31	0.000	97.12
3	20	0.497	0.13	0.70	[0.438, 0.552]			
6	4	0.732	0.69	0.75	[0.699, 0.762]			
7	5	0.702	0.64	0.71	[0.686, 0.717]			
8	10	0.727	0.62	0.83	[0.680, 0.767]			
9	20	0.753	0.58	0.90	[0.702, 0.796]			
14	32	0.799	0.62	0.88	[0.772, 0.823]			
Response Format						6.94	0.031	97.17
5-point	4	0.605	0.501	0.720	[0.508, 0.688]			
6-point	72	0.721	0.130	0.900	[0.687, 0.752]			
7-point	12	0.722	0.480	0.880	[0.634, 0.791]			
Language						106.87	0.000	97.33
English	54	0.763	0.29	0.90	[0.726, 0.795]			
Italian	6	0.795	0.72	0.84	[0.753, 0.831]			
Japanese	3	0.739	0.68	0.79	[0.674, 0.793]			
Portuguese	4	0.598	0.37	0.84	[0.427, 0.727]			
Spanish	14	0.701	0.62	0.78	[0.670, 0.730]			
Swedish	5	0.494	0.22	0.53	[0.438, 0.547]			

Note. *k* = number of reliability coefficients; *Min* = lowest reliability coefficient*; Max.* = highest reliability coefficient; *95% CI [Lb and Ub]* = lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; *Q* = heterogeneity statistic; I^2 = heterogeneity index

Table 5.
Moderator Analyses for Alpha Coefficients Across All Studies Adopting PWB Environmental Mastery Subscale

Covariate	*k*	*Mean*	*Min.*	*Max.*	*95%CI [Lb, Ub]*	*Q*	*p*	I^2
Age Category						3.77	0.288	96.42
Adolescents	11	0.648	0.28	0.88	[0.520, 0.748]			
Emerging Adults	34	0.748	0.38	0.90	[0.702, 0.787]			
Adults	47	0.737	0.42	0.92	[0.706, 0.766]			
Older Adults	9	0.687	0.00	0.89	[0.541, 0.793]			
Gender						2.68	0.262	96.33
Female	11	0.780	0.57	0.89	[0.708, 0.836]			
Male	3	0.677	0.37	0.84	[0.369, 0.851]			
Both	102	0.720	0.00	0.92	[0.695, 0.743]			
Number of Items						76.66	0.000	96.72
3	22	0.575	0.00	0.77	[0.514, 0.629]			
5	9	0.672	0.53	0.76	[0.634, 0.707]			
6	9	0.647	0.38	0.81	[0.572, 0.711]			
7	5	0.741	0.67	0.82	[0.690, 0.785]			
9	17	0.773	0.65	0.89	[0.740, 0.802]			
14	31	0.822	0.45	0.92	[0.784, 0.853]			
Response Format						14.67	0.002	96.62
4-point	3	0.671	0.65	0.70	[0.631, 0.707]			
5-point	5	0.572	0.45	0.74	[0.448, 0.675]			
6-point	71	0.739	0.28	0.92	[0.708, 0.767]			
7-point	13	0.701	0.26	0.90	[0.602, 0.779]			
Language						44.53	0.000	96.93
English	53	0.783	0.00	0.92	[0.749, 0.813]			
Italian	6	0.805	0.69	0.86	[0.744, 0.853]			
Japanese	3	0.582	0.45	0.66	[0.425, 0.705]			
Portuguese	4	0.557	0.28	0.81	[0.350, 0.712]			
Spanish	18	0.655	0.37	0.76	[0.617, 0.691]			

Swedish	5	0.634	0.42	0.73	[0.540, 0.712]

Note. *k* = number of reliability coefficients; *Min* = lowest reliability coefficient*; Max.* = highest reliability coefficient; *95% CI [Lb, Ub]* = lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; *Q* = heterogeneity statistic; I^2 = heterogeneity index

Table 6.
Moderator Analyses for Alpha Coefficients Across All Studies Adopting PWB Personal Growth Subscale

Covariate	*k*	*Mean*	*Min.*	*Max.*	*95%CI [Lb, Ub]*	*Q*	*p*	I^2
Age Category						4.65	0.199	96.59
Adolescents	10	0.649	0.18	0.81	[0.555, 0.727]			
Emerging Adults	34	0.740	0.46	0.88	[0.704, 0.772]			
Adults	52	0.737	0.40	0.91	[0.705, 0.766]			
Older Adults	12	0.731	0.36	0.87	[0.619, 0.814]			
Gender						0.36	0.535	96.68
Female	9	0.741	0.18	0.91	[0.696, 0.780]			
Both	108	0.726	0.61	0.87	[0.700, 0.749]			
Number of Items						100.02	0.000	97.05
3	21	0.547	0.18	0.86	[0.484, 0.604]			
4	4	0.657	0.57	0.71	[0.588, 0.716]			
5	4	0.719	0.68	0.80	[0.676, 0.757]			
7	15	0.726	0.52	0.79	[0.702, 0.747]			
9	18	0.765	0.60	0.85	[0.730, 0.796]			
14	37	0.809	0.66	0.91	[0.787, 0.830]			
Response Format						6.25	0.100	96.78
4-point	3	0.640	0.57	0.76	[0.530, 0.728]			
5-point	4	0.628	0.37	0.73	[0.454, 0.755]			
6-point	76	0.737	0.18	0.91	[0.706, 0.765]			
7-point	13	0.705	0.45	0.90	[0.620, 0.773]			
Language						35.42	0.000	97.07
English	59	0.768	0.36	0.91	[0.735, 0.797]			
Italian	6	0.786	0.69	0.87	[0.716, 0.840]			
Japanese	3	0.707	0.66	0.74	[0.648, 0.757]			
Portuguese	5	0.682	0.37	0.84	[0.496, 0.808]			
Spanish	16	0.674	0.49	0.74	[0.647, 0.700]			
Swedish	5	0.627	0.49	0.66	[0.579, 0.671]			

Note. *k* = number of reliability coefficients; *Min* = lowest reliability coefficient*; Max.* = highest reliability coefficient; *95% CI [Lb, Ub]* = lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; *Q* = heterogeneity statistic; I^2 = heterogeneity index

Table 7.
Moderator Analyses for Alpha Coefficients across All Studies Adopting Positive Relations

Covariate	*k*	*Mean*	*Min.*	*Max.*	*95%CI [Lb, Ub]*	*Q*	*p*	I^2
Age Category						7.46	0.060	97.05
Adolescents	11	0.708	0.23	0.90	[0.590, 0.797]			
Emerging Adults	34	0.810	0.45	0.90	[0.779, 0.836]			
Adults	51	0.767	0.51	0.93	[0.738, 0.794]			
Older Adults	10	0.743	0.48	0.89	[0.633, 0.824]			
Gender						3.69	0.158	96.93
Female	11	0.816	0.61	0.89	[0.765, 0.857]			
Male	4	0.814	0.71	0.91	[0.691, 0.892]			
Both	105	0.766	0.23	0.93	[0.743, 0.788]			
Number of Items						168.21	0.000	97.16
3	19	0.555	0.23	0.81	[0.505, 0.600]			
5	7	0.763	0.63	0.84	[0.689, 0.820]			
6	9	0.780	0.69	0.85	[0.740, 0.814]			
7	7	0.778	0.74	0.83	[0.764, 0.790]			
9	18	0.790	0.65	0.88	[0.760, 0.816]			
14	37	0.846	0.69	0.93	[0.825, 0.865]			
Response Format						13.09	0.004	97.26
4-point	3	0.672	0.63	0.74	[0.610, 0.726]			
5-point	7	0.782	0.42	0.91	[0.613, 0.883]			
6-point	75	0.777	0.23	0.93	[0.749, 0.802]			
7-point	14	0.784	0.51	0.90	[0.704, 0.844]			
Language						47.70	0.000	97.22
English	59	0.812	0.48	0.93	[0.783, 0.837]			
Italian	7	0.825	0.77	0.87	[0.787, 0.857]			
Japanese	3	0.778	0.70	0.85	[0.690, 0.843]			
Portuguese	4	0.602	0.42	0.83	[0.449, 0.721]			
Spanish	17	0.752	0.59	0.84	[0.716, 0.784]			

Swedish	5	0.554	0.30	0.65	[0.430, 0.657]

Note. *k* = number of reliability coefficients; *Min* = lowest reliability coefficient*; Max.* = highest reliability coefficient; *95% CI [Lb, Ub]* = lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; *Q* = heterogeneity statistic; I^2 = heterogeneity index

Table 8.
Moderator Analyses for Alpha Coefficients Across All Studies Adopting PWB Purpose in Life Subscale

Covariate	*k*	*Mean*	*Min.*	*Max.*	*95%CI [Lb, Ub]*	*Q*	*p*	I^2
Age Category						8.19	0.042	98.07
Adolescents	11	0.674	0.23	0.90	[0.536, 0.777]			
Emerging Adults	35	0.788	0.45	0.90	[0.757, 0.816]			
Adults	54	0.738	0.51	0.93	[0.688, 0.782]			
Older Adults	20	0.727	0.48	0.89	[0.667, 0.777]			
Gender						3.17	0.205	98.17
Female	12	0.750	-0.35	0.90	[0.638, 0.830]			
Male	3	0.712	0.68	0.73	[0.679, 0.743]			
Both	120	0.749	0.18	0.97	[0.721, 0.774]			
Number of Items						99.31	0.000	98.34
3	20	0.418	-0.35	0.88	[0.313, 0.513]			
5	7	0.768	0.66	0.84	[0.702, 0.821]			
6	10	0.775	0.65	0.88	[0.741, 0.806]			
7	13	0.744	0.65	0.88	[0.716, 0.769]			
9	22	0.754	0.58	0.88	[0.721, 0.783]			
14	39	0.841	0.57	0.97	[0.809, 0.868]			
Response Format						8.72	0.033	98.27
4-point	3	0.697	0.65	0.73	[0.643, 0.743]			
5-point	5	0.598	0.38	0.77	[0.448, 0.714]			
6-point	86	0.751	-0.35	0.93	[0.719, 0.780]			
7-point	15	0.737	0.21	0.90	[0.607, 0.828]			
Language						542.64	0.000	98.45
English	70	0.786	0.48	0.93	[0.756, 0.813]			
Italian	6	0.813	0.77	0.87	[0.760, 0.855]			
Japanese	3	0.730	0.70	0.85	[0.562, 0.840]			
Portuguese	4	0.554	0.42	0.83	[0.326, 0.721]			
Spanish	15	0.764	0.59	0.84	[0.720, 0.802]			

Swedish	5	0.254	0.30	0.65	[0.217, 0.291]

Note. *k* = number of reliability coefficients; *Min* = lowest reliability coefficient; *Max.* = highest reliability coefficient; *95% CI [Lb, Ub]* = lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; *Q* = heterogeneity statistic; I^2 = heterogeneity index

Table 9.

Moderator Analyses for Alpha Coefficients Across All Studies Adopting PWB Self-Acceptance Subscale

Covariate	*k*	*Mean*	*Min.*	*Max.*	*95%CI [Lb, Ub]*	*Q*	*p*	I^2
Age Category						8.66	0.034	97.38
Adolescents	10	0.779	0.48	0.93	[0.684, 0.848]			
Emerging Adults	37	0.837	0.61	0.93	[0.811, 0.860]			
Adults	50	0.795	0.17	0.95	[0.764, 0.822]			
Older Adults	9	0.743	0.52	0.92	[0.636, 0.821]			
Gender						4.35	0.114	97.35
Female	9	0.851	0.74	0.92	[0.803, 0.889]			
Male	5	0.802	0.65	0.89	[0.731, 0.855]			
Both	104	0.795	0.17	0.95	[0.772, 0.816]			
Number of Items						131.57	0.000	97.60
3	22	0.628	0.17	0.86	[0.577, 0.673]			
4	7	0.798	0.72	0.89	[0.728, 0.851]			
6	9	0.767	0.61	0.88	[0.701, 0.821]			
7	7	0.824	0.73	0.91	[0.797, 0.848]			
9	18	0.800	0.45	0.91	[0.741, 0.847]			
14	35	0.870	0.70	0.95	[0.853, 0.886]			
Response Format						10.35	0.016	97.49
4-point	3	0.743	0.71	0.77	[0.699, 0.781]			
5-point	6	0.709	0.36	0.80	[0.563, 0.812]			
6-point	73	0.810	0.48	0.95	[0.784, 0.833]			
7-point	13	0.799	0.17	0.92	[0.697, 0.870]			
Language						81.61	0.000	97.76
English	55	0.838	0.17	0.95	[0.809, 0.863]			
Italian	7	0.875	0.83	0.90	[0.850, 0.895]			
Japanese	3	0.813	0.78	0.83	[0.788, 0.836]			
Portuguese	5	0.662	0.36	0.83	[0.467, 0.796]			
Spanish	15	0.774	0.64	0.89	[0.727, 0.814]			

Swedish	5	0.700	0.56	0.76	[0.662, 0.734]

Note. k = number of reliability coefficients; *Min* = lowest reliability coefficient; *Max.* = highest reliability coefficient; *95% CI [Lb, Ub]* = lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; Q = heterogeneity statistic; I^2 = heterogeneity index

Figure 1.

Flowchart Describing the Search Strategy to Select Studies for Inclusion in Meta-Analysis